Efficient List-Decodable Regression using Batches

Abhimanyu Das [1] **Ayush Jain** [2] **Weihao Kong** [1] **Rajat Sen** [1]

Abstract

We demonstrate the use of batches in studying list-decodable linear regression, in which only $\alpha \in (0, 1]$ fraction of batches contain genuine samples from a common distribution and the rest can contain arbitrary or even adversarial samples. When genuine batches have $\geq \tilde{\Omega}(1/\alpha)$ samples each, our algorithm can efficiently find a small list of potential regression parameters, with a high probability that one of them is close to the true parameter. This is the first polynomial time algorithm for list-decodable linear regression, and its sample complexity scales nearly linearly with the dimension of the covariates. The polynomial time algorithm is made possible by the batch structure and may not be feasible without it, as suggested by a recent Statistical Query lower bound (Diakonikolas et al., 2021b).

1. Introduction

Linear regression is one of the most fundamental tasks in supervised learning with applications in various sciences and industries (McDonald, 2009; Dielman, 2001). In the standard linear regression setup, one is given m samples (x_i, y_i) such that $y_i = \langle w^*, x_i \rangle + n_i$ where n_i is the observation noise with bounded variance and the covariates $x_i \in \mathbb{R}^d$ are drawn i.i.d from some fixed distribution. For this setup, the commonly used least-squares estimator that minimizes the square loss $\sum_i (y_i - \langle w, x_i \rangle)^2$, provides a good estimate of the unknown regression vector w^*.

In many applications, some samples are inadvertently or maliciously corrupted, for example, due to mislabeling or measurement errors, or data poisoning attacks. For instance, such corruptions are commonplace in biology (Rosenberg et al., 2002; Paschou et al., 2010) and machine learning security (Barreno et al., 2010; Biggio et al., 2012). Even a small number of corrupt samples in the data can cause the least-squares estimator to fail catastrophically. Classical robust estimators have been proposed in (Huber, 2011; Rousseeuw, 1991) but they suffer from exponential runtime. Recent works (Lai et al., 2016; Diakonikolas et al., 2019a; 2017) have derived efficient algorithms for robust mean estimation with provable guarantees even when a small fraction of the data can be corrupt or adversarial. These works have inspired the efficient algorithms for robust regression (Prasad et al., 2018; Diakonikolas et al., 2019b;c; Pensia et al., 2020) under the same corruption model. (Cherapanamjeri et al., 2020a; Jambulapati et al., 2021) have obtained robust regression algorithms with near-optimal run time and sample complexity.

In this paper, we are interested in the setting where a small fraction α, potentially even less than half, of the data is considered inlier, and the majority of the data may be influenced by factors such as adversarial manipulation, corruption, bias, or being drawn from a diverse distribution. This setting also encompasses the problem of learning a mixture of regressions (Jordan & Jacobs, 1994; Zhong et al., 2016; Kong et al., 2020b; Pal et al., 2022) because any solution of the former immediately yields a solution to the latter by setting α to be the proportion of the data from the smallest mixture component.

However, it is information-theoretically impossible to output a single accurate estimate of regression parameter when $\alpha < 1/2$. Instead, it may be possible to generate a short list of estimates such that at least one of them is accurate. This relaxed notion of learning is known as *list-decodable learning* and is useful since a learner can identify a single accurate estimate from the list given a small number of reliable samples.

For high dimensional mean estimation, Charikar et al. (2017) derived the first polynomial time algorithm for list decodable setting. List-decodable linear regression has been studied in (Karmalkar et al., 2019; Raghavendra & Yau, 2020) yielding algorithms with runtime and sample complexity of $O(d^{\mathrm{poly}(1/\alpha)})$. In contrast to list-decodable mean estimation, recent work (Diakonikolas et al., 2021b) has shown that a sub-exponential runtime and sample complexity might be impossible for linear regression. These prior results may lead to a pessimistic conclusion for

[1]Google Research [2]UC San Diego. The majority of this work was completed while Ayush Jain was an intern at Google Research. Correspondence to: Ayush Jain <ayjain@ucsd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

obtaining practical algorithms for the fundamental learning paradigm of linear regression when less than half of the data may be inlier or genuine.

However, our work demonstrates that it can be overcome in various real-world applications such as federated learning (Wang et al., 2021), learning from multiple sensors (Wax & Ziskind, 1989), and crowd-sourcing (Steinhardt et al., 2016). In these and many other applications individual data sources often provide multiple samples. We refer to a collection of samples from a single source as a *batch*. If a fraction α of the sources follow the underlying distribution we aim to learn, then α fraction of the batches will contain independent samples from that distribution, while the remaining batches may contain arbitrary samples.

When each batch contains $\tilde{\Omega}(d)$ samples then one can get the estimate of the regression vectors for each batch. However, typically in modern applications the dimension of the data is high and only a moderate number of samples are available per batch (Grottke et al., 2015; Park & Tuzhilin, 2008; Kong et al., 2020b). As we show in this paper, for any $\alpha \in (0, 1]$, as long as the number of samples provided by each genuine source is more than a small threshold of $\tilde{\Omega}(1/\alpha)$, we can use the grouping of samples in batches to develop a polynomial-time algorithm.

The batch setting has a natural advantage in the context of list-decodable learning. When there are multiple possible inlier distributions for the data sources, the list will include regression vectors for all distributions that underlie more than α fraction of sources. To determine the best-fitting solution for a specific source from the short list generated by the list-decodable algorithm, a small hold-out portion of the batch provided by that source can be used. This post hoc identification of the best weight for a source/batch is naturally not feasible in the single sample setting.

This motivates the problem of list-decodable linear regression using batches. Formally, there are m batches. Each batch has a collection of $\geq n$ regression samples which can either all come from a global regression model with true weight w^* (good batch) and noise variance σ^2 or are arbitrarily corrupted (adversarial batch). The task is to output a small list of regression vectors at least one of which is approximately correct given that only α fraction of the batches are good. It is important to highlight that in this scenario, any algorithm aiming to provide reasonable estimation guarantees must return a list of estimates. This is because the formulation allows for data to stem from $\Theta(1/\alpha)$ different distributions, each of which generates at least α fraction of the batches. The regression parameters for each of these distributions can vary arbitrarily. Without any method to identify the genuine distribution among these $\Theta(1/\alpha)$ possibilities, any algorithm providing a single estimate of the regression parameter would fail to offer a meaningful estimation guarantee.

Our main result is the following theorem:

Theorem 1.1 (Informal)**.** *For any $\alpha \in (0, 1]$, there exists a polynomial time algorithm for list-decodable regression, that uses $m = \tilde{O}_{n,\alpha}(d)$ batches each of size $n = \tilde{\Omega}(1/\alpha)$, and outputs $O(1/\alpha^2)$ weights such that with high probability at least one of them, $\tilde{w}$, satisfies $\|\tilde{w} - w^*\|_2 = \tilde{\mathcal{O}}(\sigma/\sqrt{n\alpha})$.*

We formally state the problem in Section 2, introduce necessary notation in Section 3, and present our main result in Section 4. In Section 5, we describe the main ideas behind our algorithm and provide a comprehensive overview of our technical contributions. We present our algorithm and prove its performance guarantee in Section 6. We provide a detailed discussion of related work in Appendix A.

2. Problem formulation

We have m sources. Of these m sources at least α-fraction of the sources are genuine and provide $\geq n$ i.i.d. samples from a common distribution. The remaining sources may provide arbitrary data. Since, we can use only the first n samples from each source and ignore the rest, hence, w.l.o.g. we assume that each source provides exactly n samples. We will refer to the collection of all samples from a single source as a *batch*.

To formalize the setting, let B be a collection of m batches. Each batch $b \in B$ in this collection, has n samples $\{(x_i^b, y_i^b)\}_{i=1}^n$, where $x_i^b \in \mathbb{R}^d$ and $y_i^b \in \mathbb{R}$.

Among these batches B, there is a sub-collection G of *good batches* such that for each $b \in G$ and $i \in [n]$ samples (x_i^b, y_i^b) are generated independently from a common distribution $\mathcal{D}$ and the size of this sub-collection is $|G| \geq \alpha|B|$. The remaining batches $B \setminus G$ are *adversarial batches* and have arbitrary samples that may be selected by an adversary depending on good batches.

Next, we describe the assumption of distribution $\mathcal{D}$. We require the same set of general assumptions on the distribution, as in the recent work (Cherapanamjeri et al., 2020a), which focuses on the case when $n = 1$ and $1 - \alpha$ is small, that is when all but a small fraction of data is genuine.

Distribution Assumptions. For an unknown d-dimensional vector w^*, the *sample noises* n_i^b, the *covariates* x_i^b and the outputs y_i^b are random variables that are related as $y_i^b = x_i^b \cdot w^* + n_i^b$. Let $\Sigma = \mathbb{E}_{\mathcal{D}}[x_i^b (x_i^b)^\intercal]$. For scaling purposes, we assume $\|\Sigma\| = 1$. We have the following general assumptions.

1. x_i^b is $L4$-$L2$ hypercontractive, that is for some $C \geq 1$ and all vectors u, $\mathbb{E}_{\mathcal{D}}[(x_i^b \cdot u)^4] \leq C\, \mathbb{E}_{\mathcal{D}}[(x_i^b \cdot u)^2]^2$.

2. For some constant $C_1 > 0$, $\|x_i^b\| \le C_1\sqrt{d}$ a.s.

3. The condition number of Σ is at most C_3, that is for each unit vector u, we have $u^\intercal \Sigma u \ge \frac{\|\Sigma\|}{C_3} = \frac{1}{C_3}$.

4. Sample noise n_i^b is independent of x_i^b, has zero mean $\mathbb{E}_{\mathcal{D}}[n_i^b] = 0$, and bounded covariance $\mathbb{E}_{\mathcal{D}}[(n_i^b)^2] \le \sigma^2$.

5. The distribution of noise n_i^b is symmetric around 0.

We note that the assumptions 1,3, and 4 are standard in heavy-tailed linear regression (Cherapanamjeri et al., 2020b; Lecué & Mendelson, 2016). Assumptions 2 and 5, on the other hand, are introduced solely for the ease of presentation and we discuss in Appendix G that these two assumptions can be eliminated without any impact on our results.

3. Notation

We use h^b to denote a function over batches. For a function h^b, we use $\mathbb{E}_{\mathcal{D}}[h^b]$ and $\mathrm{Cov}_{\mathcal{D}}(h^b)$ to denote the expected value and covariance of h^b for a random batch b of n independent samples from $\mathcal{D}$.[1]

Next, we define the expectation and covariance w.r.t. the collection of batches B. When batches are chosen uniformly from a sub-collection $B' \subseteq B$, the expected value and co-variance of a function h^b are denoted as $\mathbb{E}_{B'}[h^b] = \sum_{b\in B'} \frac{1}{|B'|} h^b$ and $\mathrm{Cov}_{B'}(h^b) = \sum_{b\in B'} \frac{1}{|B'|}(h^b - \mathbb{E}_{B'}[h^b])(h^b - \mathbb{E}_{B'}[h^b])^\intercal$, respectively.

To allow for more general samplings, the definition is extended to use a weight vector. A *weight vector*, denoted by β, is a collection of weights, β^b, for each batch, $b \in B$, such that β^b is between 0 and 1. The *total weight* of the vector is represented by $\beta^B = \sum_{b\in B} \beta^b$. It can be helpful to think of β as a soft cluster of batches, with its components denoting the membership weight of batches in the cluster.

When defining expectation or covariance of a function w.r.t. a weight vector β, the probability of sampling a batch, b, is $\frac{\beta^b}{\beta^B}$. The expectation of a function, h^b, over batches, when using a weight vector β, is represented by $\mathbb{E}_\beta[h^b] := \sum_{b\in B} \frac{\beta^b}{\beta^B} h^b$, and the covariance is represented by $\mathrm{Cov}_\beta(h^b) := \sum_{b\in B} \frac{\beta^b}{\beta^B}(h^b - \mathbb{E}_\beta[h^b])(h^b - \mathbb{E}_\beta[h^b])^\intercal$.

For weight vector β, the *weight of all batches of a subset* B' is denoted as $\beta^{B'} := \sum_{b\in B'} \beta^b$.

We use $f(x) = \tilde{\mathcal{O}}(g(x))$ as a shorthand for $f(x) = \mathcal{O}(g(x)\log^k x)$, where k is some integer, and $f(x) = \mathcal{O}_y(g(x))$ implies that if y is bounded then $f(x) = \mathcal{O}(g(x))$.

[1] With slight abuse of notation, instead of $h(b)$, we use h^b to denote function over batches. Note that h^b may be a function of some or all the samples in the batch b.

Throughout the paper, we use the notation c_i, with $i \ge 1$, to represent universal constants.

4. Main Results

Recently there has been a significant interest in the problem of list decodable linear regression. The prior works considered only the non-batch setting. The sample and time complexity of algorithm in (Karmalkar et al., 2019; Raghavendra & Yau, 2020) are $d^{\mathcal{O}(1/\alpha^4)}$ and $d^{\mathcal{O}(1/\alpha^8)}$, respectively. (Raghavendra & Yau, 2020) achieves an error $\mathcal{O}(\sigma/\alpha^{3/2})$ with a list of size $(1/\alpha)^{\mathcal{O}(\log(1/\alpha))}$, and (Karmalkar et al., 2019) obtains an error guarantee $\mathcal{O}(\sigma/\alpha)$ with a list of size $\mathcal{O}(1/\alpha)$.

(Diakonikolas et al., 2021b) improved the sample complexity. For Gaussian noise and covariates distributed according to standard Gaussian, they gave an information-theoretic algorithm that uses $\mathcal{O}(d/\alpha^3)$ samples and estimates w to an accuracy $\mathcal{O}(\sigma\sqrt{\log(1/\alpha)}/\alpha)$ using a list of size $\mathcal{O}(1/\alpha)$. They also showed that no algorithm, even with infinite samples, can achieve an error $\ll \sigma/\alpha\sqrt{\log(1/\alpha)}$ with a Poly$(1/\alpha)$ size list.

As these works considered the non-batch setting, they do not obtain a polynomial time algorithm for this problem, which may in fact be impossible (Diakonikolas et al., 2021b).

Our main result shows that using batches one can achieve a polynomial time algorithm for this setting, moreover, the algorithm requires only $\tilde{\mathcal{O}}_{n,\alpha}(d)$ genuine samples.

Theorem 4.1. *For any* $0 < \alpha < 1$, $n \ge \Theta(\frac{{C_3}^2 C^2 \log^2(2/\alpha)}{\alpha})$ *and* $|G| = \Omega_C\big(dn^2\log(d)\big)$, *Algorithm 1 runs in time poly*$(|G|, \alpha, d, n)$ *and returns a list* M *of size at most* $4/\alpha^2$ *such that with probability* $\ge 1 - 4/d^2$,

$$\min_{w\in L} \|w - w^*\| \le \mathcal{O}\Big(\frac{C_3 C\log(2/\alpha)}{\sqrt{n\alpha}}\sigma\Big).$$

Interestingly, for $n = \tilde{\Omega}(1/\alpha)$, the estimation error of our polynomial algorithm has a better dependence on α than the best possible $\sigma/\alpha\sqrt{\log(1/\alpha)}$ (Diakonikolas et al., 2021b) by any algorithm (even with infinite resources) in the non-batch setting (i.e. $n = 1$).

We restate the above result as the following corollary, which for a given ϵ, d and α characterizes the number of good batches $|G|$ and n required by Algorithm 1 to achieve an estimation error $\mathcal{O}(\epsilon\sigma)$.

Corollary 4.2. *For any* $0 < \alpha < 1$, $0 \le \epsilon \le 1$, $n_{\min} = \Theta_{C_3,C}(\frac{\log^2(2/\alpha)}{\alpha\epsilon^2})$, $n \ge n_{\min}$, *and* $|G| = \Omega_C\big(dn_{\min}^2\log(d)\big)$, *Algorithm 1 runs in time poly*(α, d, ϵ) *and returns a list* M *of size at most* $4/\alpha^2$ *such that with probability* $\ge 1 - 4/d^2$,

$$\min_{w\in L} \|w - w^*\| \le \mathcal{O}(\epsilon\sigma).$$

For $\epsilon = \Theta(1)$ in the above corollary, we get $n = \tilde{\Omega}(\frac{1}{\alpha})$ and $|G| = \tilde{\Omega}_C\big(d\log(d)/\alpha^2\big)$.

Remark 4.1. As discussed earlier, for the case where a majority of data is genuine, i.e. $\alpha > 1/2$, polynomial time algorithms have been developed in prior works (Prasad et al., 2018; Diakonikolas et al., 2019b; Cherapanamjeri et al., 2020b) to estimate the regression parameter even in a non-batch setting. Since the majority of data is genuine, these algorithms can return a single estimate of the regression parameter instead of a list. In particular, the algorithm in (Cherapanamjeri et al., 2020b) requires $\mathcal{O}(d/(1-\alpha)^2)$ genuine samples, and estimates the regression parameter w^* to an ℓ_2 distance of $\mathcal{O}(C_3\sqrt{(1-\alpha)}\sigma)$ for any $1-\alpha = \mathcal{O}(\frac{1}{C_3{}^2})$, where C_3 is the condition number of the covariance matrix Σ of the covariates. A lower bound of $\Omega(\sqrt{(1-\alpha)}\sigma)$ is also known for the non-batch setting. We note that the algorithm in (Cherapanamjeri et al., 2020b) for the case $\alpha > 1/2$, can easily be extended to the batch setting, where by using batch gradients instead of sample gradients in their algorithm, the regression parameter w^* can be estimated to a much smaller ℓ_2 distance of $\mathcal{O}(C_3\sqrt{(1-\alpha)}\sigma/\sqrt{n})$.

5. Technical Overview

This section presents the main ideas behind our algorithm.

For a given batch b from B, the square loss of its ith sample at point w in the parameter space is represented by $f_i^b(w) := (w \cdot x_i^b - y_i^b)^2/2$.

If all batches in B had samples generated from $\mathcal{D}$ then the minimizer of the average loss across all batches, represented by $\mathbb{E}_B[f_i^b(w)]$, would converge to the optimal solution w^*. However, the presence of even a single outlier sample can cause this method to fail. In our setting, a majority of batches may contain potentially outlier samples.

The gradient of the loss function $f_i^b(w)$ is $\nabla f_i^b(w) = (w \cdot x_i^b - y_i^b) \cdot x_i^b$. For good batches, which has i.i.d. samples from distribution $\mathcal{D}$, the expected value of this gradient is $\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w)] = \Sigma(w - w^*)$.

When $|G|$ is sufficiently large, then

$$\|\mathbb{E}_G[\nabla f_i^b(w)]\| = \Big\|\tfrac{1}{|G|n}\textstyle\sum_{b\in G}\sum_{i\in[n]}\nabla f_i^b(w)\Big\| \approx \|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w)]\| = \|\Sigma(w-w^*)\|. \quad (1)$$

Suppose $\tilde{w}$ is a stationary point of all samples, i.e. $\mathbb{E}_B[\nabla f_i^b(\tilde{w})] = 0$. If $\tilde{w}$ is far from w^*, then the above equation implies that the mean of gradients good samples will be large. Then norm of the co-variance of the sample gradients at $\tilde{w}$ will be at least

$$\|\mathrm{Cov}_B[\nabla f_i^b(\tilde{w})]\| \geq \tfrac{|G|}{|B|}\|\mathbb{E}_G[\nabla f_i^b(\tilde{w})] - \mathbb{E}_B[\nabla f_i^b(\tilde{w})]\|^2 = \alpha\|\mathbb{E}_G[\nabla f_i^b(\tilde{w})]\|^2 \overset{(a)}{\approx} \alpha\|\Sigma(\tilde{w}-w^*)\|^2. \quad (2)$$

When the co-variance of good sample points is much smaller than the overall co-variance of all samples it is possible to iteratively divide or filter samples in two (possibly overlapping) clusters such that one of the clusters is "cleaner" than the original (Steinhardt et al., 2016; Diakonikolas et al., 2020b). Hence, if we had $\|\mathrm{Cov}_B[\nabla f_i^b(\tilde{w})]\| \gg \|\mathrm{Cov}_G[\nabla f_i^b(\tilde{w})]\|$ then we could have obtained a "cleaner version" of B, that had a higher fraction of good batches.

For batch $b \in G$ the norm of co-variance of gradients (of a single sample) is $\|\mathrm{Cov}_{\mathcal{D}}[\nabla f_i^b(\tilde{w})]\| = \Theta(\sigma^2 + \|\Sigma(\tilde{w}-w^*)\|^2)$ (using L_4-L_2 hypercontractivity). Even if we had $\|\mathrm{Cov}_G[\nabla f_i^b(\tilde{w})]\| \approx \|\mathrm{Cov}_{\mathcal{D}}[\nabla f_i^b(\tilde{w})]\|$, it does not guarantee $|\mathrm{Cov}_B[\nabla f_i^b(\tilde{w})]| \gg |\mathrm{Cov}_G[\nabla f_i^b(\tilde{w})]|$, as $\alpha\|\Sigma(\tilde{w}-w^*)\|^2 \ll \sigma^2 + \|\Sigma(\tilde{w}-w^*)\|^2$, regardless of how large the difference between the stationary point w^* for the distribution $\mathcal{D}$ and the stationary point $\tilde{w}$ for all samples is. We will now see that focusing on batch gradients rather than single sample gradients can alleviate this problem.

5.1. How Batches Help

In the preceding approach, we didn't leverage the batch structure. In fact, the SQ lower bound in (Diakonikolas et al., 2021b) suggests that it may be impossible to achieve a polynomial-time algorithm for the non-batch setting.

To take the advantage of the batch structure instead of considering the loss function and its gradient for each sample individually, we consider the loss of a batch and the gradient of the batch loss. The loss function of a batch b is $f^b(w) = \frac{1}{n}\sum_{i=1}^n f_i^b(w)$ i.e the average of the loss function in its samples. From the linearity of differentiation, the gradient of the batch loss function is $\nabla f^b(w) = \frac{1}{n}\sum_{i=1}^n \nabla f_i^b(w)$.

Then from the linearity of expectation, $\|\mathbb{E}_G[\nabla f_i^b(w)]\| = \|\mathbb{E}_G[\nabla f^b(w)]\|$ for any w. However, averaging over n samples reduces the co-variance by a factor n, therefore, $\mathrm{Cov}_{\mathcal{D}}[\nabla f^b(w)] = \mathrm{Cov}_{\mathcal{D}}[\nabla f_i^b(w)]/n$.

For $|G|$ large enough, we will have population covariance $\|\mathrm{Cov}_G[\nabla f^b(\tilde{w})]\| \approx \|\mathrm{Cov}_{\mathcal{D}}[\nabla f^b(\tilde{w})]\|$. Further, as $\|\mathrm{Cov}_{\mathcal{D}}[\nabla f_i^b(\tilde{w})]\| = \mathcal{O}\big(\sigma^2 + \|\Sigma(\tilde{w}-w^*)\|^2\big)$, it follows

$$\|\mathrm{Cov}_G[\nabla f^b(\tilde{w})] = \mathcal{O}\Big(\tfrac{\sigma^2 + \|\Sigma(\tilde{w}-w^*)\|^2}{n}\Big).$$

If the batch size $n = \Omega(\log^2(1/\alpha)/\alpha)$ and $\tilde{w}$ is stationary point of average loss of all samples in B, then for a large value of $\|\tilde{w}-w^*\| = \Omega(\sigma\log(1/\alpha)/\sqrt{n\alpha})$, it can be shown that $\alpha\|\Sigma(\tilde{w}-w^*)\|^2 \geq \log^2(\frac{1}{\alpha})\mathcal{O}(\frac{\sigma^2+\|\Sigma(\tilde{w}-w^*)\|^2)}{n})$. Since the expectation of batch and sample gradients are the same over any batch sub-collection, using a similar argument as for Equation (2) one can show that $\|\mathrm{Cov}_B[\nabla f^b(\tilde{w})]\| \approx \alpha\|\Sigma(\tilde{w}-w^*)\|^2$, Combining this bound with the above bound gives

$$\|\mathrm{Cov}_B[\nabla f^b(\tilde{w})]\| \geq \log^2(1/\alpha)\|\mathrm{Cov}_G[\nabla f^b(\tilde{w})]\|.$$

Therefore, either the distance between the stationary point of this cluster and w^* is $\leq \mathcal{O}(\frac{\sigma \log(1/\alpha)}{\sqrt{n\alpha}})$ or the covariance of gradients for the set of all batches is much larger than that for good batches. If it is the former, then we have a good approximation of w^* and if it is the latter, we can divide B into two (possibly overlapping) clusters, where at least one of the new clusters contains a majority of good batches and has a higher proportion of good batches than the initial cluster. The same argument can be extended from B to any sub-collection of B that retains a major portion of good batches G.

To divide the clusters, we use the MULTIFILTER routine from (Diakonikolas et al., 2020b). Instead of hard clustering, this routine does soft clustering. The soft clustering produces a membership or weight vector β of length $|B|$ with each entry between $[0, 1]$ that denotes the membership weight of the corresponding batch in the cluster.

The above discussion leads to the following algorithm. We begin with the initial cluster of all batches B. We keep applying MULTIFILTER routine iteratively on the clusters (or weight vectors) until, for all the clusters, the covariance of gradients at the stationary points of the respective clusters becomes small. MULTIFILTER routine ensures that at least one of the clusters retains a major portion of good batches and it doesn't have more than $\mathcal{O}(1/\alpha^2)$ clusters at any stage.

As discussed, for a cluster that retains a major portion of good batches G, the covariance of batch gradients is small only if stationary point $\tilde{w}$ of that cluster approximates w^* with an accuracy of $\|\tilde{w} - w^*\| = \mathcal{O}(\frac{\sigma \log(1/\alpha)}{\sqrt{n\alpha}})$. Since the final set of $\mathcal{O}(1/\alpha^2)$ clusters includes at least one such cluster, the stationary point of at least one of the clusters should approximate w^* to the desired accuracy.

However, applying the MULTIFILTER routine for this purpose presents additional challenges, which we address through various technical contributions in the next section.

5.2. Clipping to Improve Sample Complexity

We would like to obtain a high probability concentration bound of $\|\text{Cov}_G[\nabla f^b(w)]\| \leq \mathcal{O}(\frac{\|w-w^*\|^2+\sigma^2}{n})$ on the empirical covariance of the batch gradients in the good batches. No such bounds for general n are known in previous literature. And even for $n = 1$, using known concentration bounds would require a large number of good batches or samples. For example, (Diakonikolas et al., 2019b) needed d^5 samples in total, and in fact, a minimum requirement of d^2 samples can be shown for such a bound to hold. (Cherapanamjeri et al., 2020a) required $O(d)$ samples (for n fixed to 1) for a related bound, but for each point w they need to ignore certain samples from the calculation of empirical covariance. These samples can be different depending on w. While such guarantees sufficed for their application where a majority of data was genuine, it is unclear if it can be extended to the list-decodable setting.

To address these challenges, we use *clipped loss* instead. For *clipping parameter* $\kappa > 0$ and any batch $b \in B$, the clipped loss of its i^{th} sample at point w is given by

$$f_i^b(w, \kappa) := \begin{cases} \frac{(w \cdot x_i^b - y_i^b)^2}{2} & \text{if } |w \cdot x_i^b - y_i^b| \leq \kappa \\ \kappa |w \cdot x_i^b - y_i^b| - \kappa^2/2 & \text{otherwise.} \end{cases}$$

We specify the choice of clipping parameter κ later. The clipped loss defined above is known as Huber's loss in literature. The gradient of this clipped loss is

$$\nabla f_i^b(w, \kappa) := \frac{(x_i^b \cdot w - y_i^b)}{|x_i^b \cdot w - y_i^b| \vee \kappa} \kappa x_i^b. \tag{3}$$

We refer to the gradient of the clipped loss above as the *clipped gradient*.

For a batch b, its *clipped loss* is simply the average of clipped loss over all its samples. *Clipped loss* for a batch b at point w is $f^b(w, \kappa) := \frac{1}{n} \sum_{i \in [n]} f_i^b(w, \kappa)$. By the linearity of gradients, the gradient of the clipped loss, or *clipped gradient*, $\nabla f^b(w, \kappa)$ is the average of clipped loss over all its samples, i.e. $\nabla f^b(w, \kappa) := \sum_{i \in [n]} \frac{1}{n} \nabla f_i^b(w, \kappa)$.

Ideal choice of Clipping parameter. When $\kappa \to \infty$, the clipped loss is the same as the squared loss, hence clipping will have no effect in reducing the number of samples required. On the other hand, if $\kappa \to 0$, the loss function is overly clipped, which can lead to the expected norm of the clipped gradient being much smaller than that of the unclipped gradients in Equation (1). Theorem C.2 shows that as long as the clipping parameter is set to $\Omega(\|w - w^*\|) + \Omega_{n\alpha}(\sigma)$, the expected norm of the clipped gradient will be $\Omega(\|w - w^*\|) - \tilde{\mathcal{O}}(\sigma/\sqrt{n\alpha})$. This means that for any point w whose distance from w^* is greater than $\tilde{\Omega}(\sigma/\sqrt{n\alpha})$, the expected norm of the clipped gradient at w is $\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w, \kappa)]\| = \Omega(\|w - w^*\|)$, which is of the same order as that of unclipped gradients in Equation (1).

Furthermore, taking advantage of clipping, in Theorem B.1 we show that for all points w, and for any clipping parameter $\kappa = \mathcal{O}(\|w - w^*\|) + \mathcal{O}_{n\alpha}(\sigma)$ the covariance of the clipped gradients satisfies $\|\text{Cov}_G[\nabla f^b(w, \kappa)]\| \leq \mathcal{O}(\frac{\|w-w^*\|^2+\sigma^2}{n})$ with only $\tilde{O}_{n,\alpha}(d)$ samples. As discussed previously, the same bound on the covariance of the un-clipped gradients would instead require $\Omega(d^2)$ samples.

From the preceding discussion, in order for both the requirements of $\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w, \kappa)]\| = \Omega(\|w - w^*\|)$ and $\|\text{Cov}_G[\nabla f^b(w, \kappa)]\| \leq \mathcal{O}(\frac{\|w-w^*\|^2+\sigma^2}{n})$ to be met using only $\tilde{O}_{n,\alpha}(d)$ samples, the clipping parameter must be set to $\kappa = \Theta(\|w - w^*\|) + \Theta_{n\alpha}(\sigma)$. This requires a constant factor approximation of $\|w - w^*\|$ to be obtained.

Additionally, when using the MULTIFILTER on a cluster, a tight approximation of $\|w - w^*\|$ is necessary to obtain a tight upper bound on $\|\text{Cov}_G[\nabla f^b(w, \kappa)]\|$, which is required by MULTIFILTER as an input parameter.

Furthermore, recall that when applying the MULTIFILTER on any cluster, we set w to a stationary point of the clipped loss for that cluster. This stationary point w will depend on the clipping parameter κ, and the appropriate range for κ depends on w, creating a cyclic dependence that we must also overcome when estimating $\|w - w^*\|$.

5.3. Estimating Parameters for Multifilter

Recall that our goal is to return a small set of (soft) clusters, such that at least one of them retains a major portion of good batches, and its stationary point closely approximates w^*. When the MULTIFILTER routine is applied to a cluster, it generates sub-clusters. Hence, the sub-clusters that originate from a cluster that has already lost a majority of good batches are not relevant for us. Therefore, we will need accurate parameter estimation only for clusters that have retained a substantial weight of good batches, and will only consider such clusters in the remaining section.

Let $v^b(w) := \frac{1}{n}\sum_{i\in[n]} |w \cdot x_i^b - y_i^b|$ denote the mean absolute loss of a batch at point w.

We developed a subroutine called FINDCLIPPINGPPARAMETER (Algorithm 2) that overcomes the cyclic dependence to find appropriate stationary point w and clipping parameter κ for a given soft cluster β. These values ensure that w is a stationary point for clipped gradients $\nabla f^b(w, \kappa)$ for batches in the cluster and κ falls in a range determined by the expected absolute loss of batches in cluster β at the stationary point w, specifically, $\kappa = \Theta(\mathbb{E}_\beta[v^b(w)]) + \Theta_{n\alpha}(\sigma)$.

For $|G| = \tilde{\Omega}(d)$, we prove that w.h.p.

$$\begin{aligned}\text{Var}_G(v^b(w)) &\le \mathbb{E}_G[(v^b(w) - \mathbb{E}_{\mathcal{D}}[v^b(w)])^2] \\ &= \mathcal{O}\Big(\tfrac{\sigma^2 + \mathbb{E}_{\mathcal{D}}[v^b(w)]^2}{n}\Big). \end{aligned} \tag{4}$$

From the above bound, it follows that for most of the good batches, $v^b(w)$ is very close to $\mathbb{E}_{\mathcal{D}}[v^b(w)]$.

Further, it can be shown that $\mathbb{E}_{\mathcal{D}}[v^b(w)] = \Theta(\|w - w^*\|) \pm \mathcal{O}(\sigma)$, where $\mathbb{E}_{\mathcal{D}}[v^b(w)]$ is expectation of $v^b(w)$ for a batch sampled from $\mathcal{D}$.

We derive a novel way that given a weight vector β estimates upper bound θ_1 on variance $\text{Var}_G(v^b(w))$. Further, the upper bound is tight enough to ensure that if $\text{Var}_\beta(v^b(w)) = \tilde{\mathcal{O}}(\theta_1)$ then for most batches b in β, $v^b(w)$ will be close to its expectation $\mathbb{E}_\beta[v^b(w)]$ over β. As the soft cluster contains a significant proportion of good batches, and since $v^b(w)$ for most of these good batches is close to $\mathbb{E}_{\mathcal{D}}[v^b(w)]$, then $\mathbb{E}_\beta[v^b(w)]$ would also be close to $\mathbb{E}_{\mathcal{D}}[v^b(w)]$. Furthermore, since $\mathbb{E}_{\mathcal{D}}[v^b(w)] = \Theta(\|w - w^*\|) \pm \mathcal{O}(\sigma)$, it follows that $\mathbb{E}_\beta[v^b(w)] = \Theta(\|w - w^*\|) \pm \mathcal{O}(\sigma)$.

Therefore, if for a certain weight vector β the variance $\text{Var}_\beta(v^b(w))$, is close to our estimated variance of good batches, θ_1, then we can ensure that $\kappa = \Theta(\|w - w^*\|) + \Theta_{n\alpha}(\sigma)$ and use $\mathbb{E}_\beta[v^b(w)]$ as an estimate for $\|w - w^*\|$.

However, if the variance $\text{Var}_\beta(v^b(w))$ for a β, is significantly greater than our estimated variance of good batches, θ_1, then we will not use the MULTIFILTER routine for gradients on that cluster. Instead, we will apply MULTIFILTER routine on this cluster w.r.t. average absolute loss $v^b(w)$. As a result, the estimation of $\|w - w^*\|$ and ensuring that κ is within the correct range, which is necessary for using the MULTIFILTER routine for gradients, is no longer relevant.

Hence, in the estimation part, we either apply MULTIFILTER routine on the cluster for average absolute loss to obtain new clusters with one of them being cleaner, or else our estimate of the parameters is in the desired range to apply MULTIFILTER routine w.r.t. the gradients.

6. Algorithm and Proof of Theorem 4.1

In subsection 6.2, a triplet (β, κ, w) is defined as nice if it meets certain criteria: β retains a substantial amount of weight among good batches, κ falls within a specific range, w is a stationary point of the clipped loss, and the covariance of gradients of the clipped loss for the cluster β is bounded at w. It is noted that any such triplet's point w is a good approximation of w^*. In Section 5.1, we provided intuition for the same without the clipping.

To identify a cluster with bounded covariance of clipped gradients, we require that the covariance of clipped gradients for G is bounded. And to estimate the correct range of κ and an upper bound on the covariance of clipped gradients for G, as described in Section 5.3, we require that the variance of mean absolute loss is bounded for the set of good batches. We formalize these requirements in the next subsection in form of two regularity conditions.

To specify the range in which the clipping parameter κ should be set, we define $\kappa_{\max}$ and $\kappa_{\min}$ in Definition B.1 in the appendix, which are functions of w, and other distribution parameters. Finally, in the last two subsections, we describe the algorithm and show that it finds a nice triplet.

6.1. Regularity Conditions.

The first condition is that for all unit vectors u, all vectors w and for all $\kappa \le \kappa_{\max}$,

$$\mathbb{E}_G\Big[\big(\nabla f^b(w, \kappa) \cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w, \kappa) \cdot u]\big)^2\Big] \le U_1,$$

where $U_1 := c_4 \frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|^2]}{n}$. The second regularity condition is that for all vectors w,

$$\mathbb{E}_G\Big[\big(v^b(w) - \mathbb{E}_{\mathcal{D}}[|w\cdot x_i^b - y_i^b|]\big)^2\Big] \le U_2,$$

where $U_2 := c_2 \frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[|w\cdot x_i^b - y_i^b|]^2}{n}$. We will repeatedly refer to the upper bounds U_1 and U_2 in the regularity conditions throughout this section.

In Section B, we show that even with a minimal number of good batches, $|G| = \tilde{\Omega}_{n,\alpha}(d)$, the two regularity conditions hold w.h.p.

As a simple consequence, the first regularity condition implies that

$$\|\mathrm{Cov}_G(\nabla f^b(w,\kappa))\| \le U_1, \tag{5}$$

and similarly, the second regularity condition implies that

$$\mathrm{Var}_G\big(v^b(w)\big) \le U_2. \tag{6}$$

We note that the expressions for U_1 and U_2 simplify to $\Theta(\sigma^2 + \|w-w^*\|^2/n)$ and the expressions for $\kappa_{\max}$ and $\kappa_{\min}$ simplify to $\Theta(\|w-w^*\| + \sigma)$ if one is not concerned with the dependence on distribution parameters C, C_3, C_p.

6.2. Nice Triplet

First, we introduce the notion of *nice* weight vector. A weight vector β is considered *nice* if the total weight assigned to all good batches by it is at least $\beta^G \ge 3|G|/4$.

We term a combination of a weight vector β, a clipping parameter κ, and an estimate w as a *triplet*. Next, we introduce the concept of a *nice* triplet.

Condition 1. A triplet (β, κ, w) is considered nice if

(a) β is a nice weight vector, i.e. $\beta^G \ge 3|G|/4$.

(b) Clipping parameter is in the range, $\kappa_{\min} \le \kappa \le \kappa_{\max}$.

(c) w is an approximate stationary point, namely mean clipped loss for weight vector β at w is at most $\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\| \le \log(2/\alpha)\sigma/8\sqrt{n\alpha}$.

(d) Covariance of the clipped gradients over β at stationary point w is at most $\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \le c_5 C^2 \log^2(\frac{2}{\alpha}) \frac{(\sigma^2 + \mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n}$, where c_5 is a positive universal constant.

According to these conditions, a triplet (β, κ, w) is nice if weight vector β is considered nice, clipping parameter κ is within the appropriate range, w is an approximate stationary point for clipped loss for this weight vector and covariance of clipped gradient over weight vector β at this point w is small. As discussed briefly at the beginning of this section, for a triplet satisfying these conditions w is a good approximation of w^*. Theorem C.1 formally shows that for any nice triplet (β, w, κ), we have $\|w - w^*\| \le \mathcal{O}(\frac{C_3 C\sigma \log(2/\alpha)}{\sqrt{n\alpha}})$. Then to prove Theorem 4.1, it is sufficient to show that the algorithm returns a small list of triplets such that at least one of them is nice.

In the next two subsections, we will describe the algorithm and demonstrate that it returns a small list of triplets, at least one of which is nice.

Algorithm 1 MAINALGORITHM

Input: Data $\{\{(x_i^b, y_i^b)\}_{i\in[n]}\}_{b\in B}$, α, C, σ.
For each $b \in B$, $\beta_{init}^b \leftarrow 1$ and $\beta_{init} \leftarrow \{\beta_{init}^b\}_{b\in B}$.
List $L \leftarrow \{\beta_{init}\}$ and $M \leftarrow \varnothing$.
while $L \neq \varnothing$ **do**
  Pick any element β in L and remove it from L.
  $a_1 = \frac{256C\sqrt{2}}{3}$ and $a_2 = \frac{a_1}{4} + 64$.
  $\kappa, w \leftarrow$ FINDCLIPPINGPPARAMETER(B, β, a_1, a_2 $\{\{(x_i^b, y_i^b)\}_{i\in[n]}\}_{b\in B}$)
  Find top approximate unit eigenvector u of $\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))$.
  For each batch $b \in B$, let $v^b = \frac{1}{n}\sum_{i\in[n]} |w\cdot x_i^b - y_i^b|$ and $\tilde{v}^b = \nabla f^b(w,\kappa)\cdot u$.
  $\theta_0 \leftarrow \inf\{v : \beta(\{b : v^b \ge v\}) \le \alpha|B|/4$ and

$$\theta_1 \leftarrow \tfrac{c_2}{n}\Big(\sigma^2 + \Big(\tfrac{8\sqrt{C}\theta_0}{7} + \tfrac{\sigma}{7}\Big)^2\Big), \tag{7}$$

$$\theta_2 \leftarrow \tfrac{c_4}{n}\Big(\sigma^2 + 16C^2\big(\mathbb{E}_\beta[v^b] + \sigma\big)^2\Big). \tag{8}$$

  if $\mathrm{Var}_{B,\beta}(v^b) > c_3 \log^2(2/\alpha)\theta_1$ **then**
    NEWWEIGHTS $\leftarrow$MULTIFILTER($B, \alpha, \beta, \{v^b\}, \theta_1$). {**Type-1 use**}
    Append each weight vector $\widetilde{\beta} \in$ NEWWEIGHTS that has total weight $\widetilde{\beta}^B \ge \alpha|B|/2$ to list L.
  else if $\mathrm{Var}_{B,\beta}(\tilde{v}^b) > c_3 \log^2(2/\alpha)\theta_2$ **then**
    NEWWEIGHTS $\leftarrow$MULTIFILTER($B, \alpha, \beta, \{\tilde{v}^b\}, \theta_2$). {**Type-2 use**}
    Append each weight vector $\widetilde{\beta} \in$ NEWWEIGHTS that has total weight $\widetilde{\beta}^B \ge \alpha|B|/2$ to list L.
  else
    Append (β, κ, w) to M.
  end if
end while
Return M

6.3. Description of the Algorithm

MAINALGORITHM starts with $L = \beta_{init}$, where the initial weight vector β_{init} assigns an equal weight of 1 to each batch in B. This initial weight vector is nice since $\beta_{init}^G = |G|$. In each iteration of the while loop, the algorithm selects one of the weight vectors β from the list L, until the list L is empty. Then, it uses the subroutine FINDCLIPPINGPPARAMETER on this weight vector β, which

returns the values of clipping parameter κ and approximate stationary point of clipped loss as w.

Next, the algorithm uses the MULTIFILTER subroutine on β. Given a weight vector and a function over batches, as well as an estimate of the variance of the function for good batches, this subroutine divides the cluster to produce new clusters, such that each of them is shorter than the original.

To apply this subroutine, the algorithm first calculates parameters θ_1 and θ_2, which are estimates of the upper bounds U_2 and U_1 in the two regularity conditions for the point w.

If the variance of the mean absolute loss at w for batches in this weight vector β is much larger than the estimate θ_1, namely $\text{Var}_\beta\big(v^b\big) \geq c_3 \log^2(2/\alpha)\theta_1$, the algorithm applies the MULTIFILTER subroutine for the function $v^b(w)$. This is referred to as a Type-1 use of this subroutine.

If instead, the variance of v^b in the weight vector is small, the algorithm defines a new function on batches, $\tilde{v}^b := \nabla f^b(w,\kappa)\cdot u$, where u is a top approximate unit eigenvector of $\text{Cov}_\beta(\nabla f^b(w,\kappa))$ such that $u^\intercal \text{Cov}_\beta(\nabla f^b(w,\kappa))u \geq 0.5\|\text{Cov}_\beta(\nabla f^b(w,\kappa))\|$. This function $\tilde{v}^b$ is a projection of clipped batch gradients along the direction in which covariance is nearly the highest. From (5), it follows that variance of this new function $\tilde{v}^b$ in good batch collection G will be bounded by U_1. If the variance of $\tilde{v}^b$ over the weight vector β is much larger than estimate θ_2 of U_1, namely $\text{Var}_\beta\big(\tilde{v}^b\big) \geq c_3 \log^2(2/\alpha)\theta_2$, then the algorithm applies MULTIFILTER subroutine for function $\tilde{v}^b(w)$. This is referred to as a Type-2 use of this subroutine.

When MULTIFILTER is applied to a weight vector, it returns a list NEWWEIGHTS of weight vectors as a result. The MAINALGORITHM appends weight vectors in NEWWEIGHTS that have total weights more than $\alpha|B|/2$ to list L and the iteration terminates. The weight vectors that have total weights less than $\alpha|B|/2$ are ignored as they can't be nice weight vectors and can not result in any nice weight vector in future iterations.

If the variances of both v^b and $\tilde{v}^b$ are small, then the iteration ends by appending (β,κ,w) to M. Next, we argue that M ends up with at least one nice triplet.

6.4. Finding Nice Triplet

We first show that Type-1 application of MULTIFILTER on a nice weight β only occurs when,

$$\text{Var}_\beta\big(v^b\big) \geq c_3 \log^2(2/\alpha)\text{Var}_G\big(v^b\big). \tag{9}$$

Recall that Type-1 application of MULTIFILTER on β takes place when $\text{Var}_\beta\big(v^b\big) \geq c_3 \log^2(2/\alpha)\theta_1$. From Equation (6), we have $\text{Var}_G\big(v^b\big) \leq U_2$ and Theorem E.1 shows that for a nice weight vector β the parameter θ_1 upper bounds U_2. Thus, Type-1 use of MULTIFILTER on a nice weight β only takes place when Equation (9) holds.

The subroutine FINDCLIPPINGPPARAMETER returns κ and w for a given weight vector β. Theorem D.1 in the Appendix D shows that these parameters w and κ satisfy:

1. w is an approximate stationary point for $\{f^b(\cdot,\kappa)\}$ w.r.t. weight vector β.
2. $\big(\frac{a_1}{2}\,\mathbb{E}_\beta[v^b(w)] \vee a_2\sigma\big) \leq \kappa \leq \big(4a_1^2\,\mathbb{E}_\beta[v^b(w)] \vee a_2\sigma\big)$, where a_1 and a_2 are input parameters of FINDCLIPPINGPPARAMETER.

The first guarantee implies that if a triplet (β,κ,w) ends in set M, then it must satisfy condition (c) for a nice triplet.

Theorem E.4 shows that if Type-1 filtering did not occur for a nice weight vector, then for this weight vector the range of κ specified in the second guarantee of subroutine FINDCLIPPINGPPARAMETER is a subset of the desired range $(\kappa_{\min},\kappa_{\max})$. Specifically, if for a nice weight vector β, $\text{Var}_\beta\big(v^b\big) \leq c_3 \log^2(2/\alpha)\theta_1$, then $\kappa \in (\kappa_{\min},\kappa_{\max})$ and

$$U_1 \leq \theta_2 \leq \tfrac{c_5}{2c_3}\,\tfrac{C^2(\sigma^2+\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n}. \tag{10}$$

Recall that a triplet (β,κ,w) ends up in M only when $\text{Var}_\beta\big(v^b\big) \leq c_3 \log^2(2/\alpha)\theta_1$ and $\text{Var}_\beta\big(\tilde{v}^b\big) \leq c_3 \log^2(2/\alpha)\theta_2$ are both satisfied.

From the above discussion, it follows that if a triplet (β,κ,w) is in M such that β is nice then $\kappa \in (\kappa_{\min},\kappa_{\max})$ and it satisfies,

$$\text{Var}_\beta\big(\tilde{v}^b\big) \leq \tfrac{c_5}{2} \log^2(2/\alpha)\,\tfrac{C^2(\sigma^2+\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n}.$$

From the definition of $\tilde{v}^b$, it follows that $\|\text{Cov}_\beta(\nabla f^b(w,\kappa))\| \leq 2\text{Var}_\beta\big(\tilde{v}^b\big)$. Therefore, for any triplet (β,κ,w) in M such that β is a nice weight vector, conditions (b) and (d) are also satisfied. This means that any such triplet is a nice triplet. Finally, it remains to be shown that M contains at least one triplet with a nice weight vector, which we do next.

Recall that Type-2 application of MULTIFILTER on a weight β only takes place when, $\text{Var}_\beta\big(\tilde{v}^b\big) \geq c_3 \log^2(2/\alpha)\theta_2$. Since for a nice β, from Equation (10), $U_1 \leq \theta_2$, from Equation (5), $\|\text{Cov}_G(\nabla f^b(w,\kappa))\| \leq U_1$, and from the definition of $\tilde{v}^b$, $\text{Var}_G\big(\tilde{v}^b\big) \leq \|\text{Cov}_G(\nabla f^b(w,\kappa))\|$. Therefore, $\theta_2 \geq \text{Var}_G\big(\tilde{v}^b\big)$, and hence Type-2 application on a nice weight β only takes place when,

$$\text{Var}_\beta\big(\tilde{v}^b\big) \geq c_3 \log^2(2/\alpha)\text{Var}_G\big(\tilde{v}^b\big). \tag{11}$$

Theorem F.2 in Appendix F states that if Equation (9) holds for all Type-1 uses and Equation (11) holds for all Type-2 uses when using subroutine MULTIFILTER on nice weight vectors, then at least one of the triplets in the final list M will include a nice weight vector. Since we have already

shown that these two equations hold, it follows that M will contain a nice triplet. The theorem also shows that the size of M is at most $4/\alpha^2$ and the total number of iterations of the while loop is at most $\mathcal{O}(|B|/\alpha^2)$, implying a small list size and a polynomial runtime for the algorithm

7. Conclusion

In summary, this paper addresses the problem of linear regression in the setting when data is presented in batches and only a small fraction of the batches contain genuine data. The paper presents a polynomial time algorithm to identify a small list containing a good approximation of the true regression parameter when genuine batches have at least $\tilde{\Omega}(1/\alpha)$ samples each. By utilizing the batch structure, the paper introduces the first polynomial-time algorithm for list decodable linear regression. Additionally, the algorithm requires a number of genuine samples that increase nearly linearly with the dimension of the covariates.

SQ lower bounds in (Diakonikolas et al., 2021b) for the non-batch setting suggests that a polynomial time algorithm is impossible with batch size 1, and the paper demonstrates that a batch size of $\geq \tilde{\Omega}(1/\alpha)$ is sufficient to obtain a polynomial time algorithm. This poses the question of what the smallest batch size required is to obtain a polynomial time algorithm, which is a promising direction for future work.

References

Acharya, J., Jain, A., Kamath, G., Suresh, A. T., and Zhang, H. Robust estimation for random graphs. In *Conference on Learning Theory*, pp. 130–166. PMLR, 2022.

Anscombe, F. J. Rejection of outliers. *Technometrics*, 2(2): 123–146, 1960.

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pp. 169–212, 2017.

Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. The security of machine learning. *Machine Learning*, 81 (2):121–148, 2010.

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. Consistent robust regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 2107–2116, 2017a.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017b.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Chaganty, A. T. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning (ICML)*, pp. 1040–1048, 2013.

Charikar, M., Steinhardt, J., and Valiant, G. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

Chen, S., Li, J., and Moitra, A. Learning structured distributions from untrusted batches: Faster and simpler. *Advances in Neural Information Processing Systems*, 33: 4512–4523, 2020a.

Chen, S., Li, J., and Song, Z. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *STOC*. https://arxiv.org/pdf/1912.07629.pdf, 2020b.

Chen, Y. and Poor, H. V. Learning mixtures of linear dynamical systems. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3507–3557. PMLR, 17–23 Jul 2022.

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. P. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pp. 727–757. PMLR, 2019.

Cherapanamjeri, Y., Aras, E., Tripuraneni, N., Jordan, M. I., Flammarion, N., and Bartlett, P. L. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020a.

Cherapanamjeri, Y., Aras, E., Tripuraneni, N., Jordan, M. I., Flammarion, N., and Bartlett, P. L. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020b.

Cherapanamjeri, Y., Mohanty, S., and Yau, M. List decodable mean estimation in nearly linear time. *arXiv preprint arXiv:2005.09796*, 2020c.

Dalalyan, A. and Thompson, P. Outlier-robust estimation of a sparse linear model using ℓ_1 -penalized huber's m-estimator. *Advances in neural information processing systems*, 32, 2019.

Diakonikolas, I. and Kane, D. M. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 184–195. IEEE, 2020.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pp. 999–1008. PMLR, 2017.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being Robust (in High Dimensions) Can Be Practical. *arXiv e-prints*, art. arXiv:1703.00893, March 2017.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2683–2702. SIAM, 2018a.

Diakonikolas, I., Kane, D. M., and Stewart, A. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1047–1060, 2018b.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019a.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pp. 1596–1606. JMLR, Inc., 2019b.

Diakonikolas, I., Kong, W., and Stewart, A. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2745–2754. SIAM, 2019c.

Diakonikolas, I., Hopkins, S. B., Kane, D., and Karmalkar, S. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020a.

Diakonikolas, I., Kane, D., and Kongsgaard, D. List-decodable mean estimation via iterative multi-filtering. *Advances in Neural Information Processing Systems*, 33: 9312–9323, 2020b.

Diakonikolas, I., Kane, D., Kongsgaard, D., Li, J., and Tian, K. List-decodable mean estimation in nearly-pca time. *Advances in Neural Information Processing Systems*, 34: 10195–10208, 2021a.

Diakonikolas, I., Kane, D., Pensia, A., Pittas, T., and Stewart, A. Statistical query lower bounds for list-decodable linear regression. *Advances in Neural Information Processing Systems*, 34:3191–3204, 2021b.

Dielman, T. E. *Applied regression analysis for business and economics*. Duxbury/Thomson Learning Pacific Grove, CA, 2001.

Dong, Y., Hopkins, S., and Li, J. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.

Gao, C. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

Grottke, M., Knoll, J., and Groß, R. How the distribution of the number of items rated per user influences the quality of recommendations. In *2015 15th International Conference on Innovations for Community Services (I4CS)*, pp. 1–8. IEEE, 2015.

Hopkins, S., Li, J., and Zhang, F. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33, 2020a.

Hopkins, S. B. and Li, J. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1021–1034, 2018.

Hopkins, S. B. et al. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2): 1193–1213, 2020b.

Huber, P. J. Robust estimation of a location parameter. *Annals Mathematics Statistics, 35*, 1964.

Huber, P. J. Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer, 2011.

Jain, A. and Orlitsky, A. Optimal robust learning of discrete distributions from batches. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pp. 4651–4660. JMLR, Inc., 2020a.

Jain, A. and Orlitsky, A. A general method for robust learning from batches. *arXiv preprint arXiv:2002.11099*, 2020b.

Jain, A. and Orlitsky, A. Robust density estimation from batches: The best things in life are (nearly) free. In *International Conference on Machine Learning*, pp. 4698–4708. PMLR, 2021.

Jambulapati, A., Li, J., and Tian, K. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33, 2020.

Jambulapati, A., Li, J., Schramm, T., and Tian, K. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34:4475–4488, 2021.

Jia, H. and Vempala, S. Robustly clustering a mixture of gaussians. *arXiv preprint arXiv:1911.11838*, 2019.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.

Karmalkar, S. and Price, E. Compressed sensing with adversarial sparse noise via l1 regression. In *2nd Symposium on Simplicity in Algorithms*, 2019.

Karmalkar, S., Klivans, A., and Kothari, P. List-decodable linear regression. *Advances in neural information processing systems*, 32, 2019.

Klivans, A., Kothari, P. K., and Meka, R. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pp. 1420–1430. PMLR, 2018.

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. *Advances in Neural Information Processing Systems*, 33, 2020a.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pp. 5394–5404. PMLR, 2020b.

Kong, W., Sen, R., Awasthi, P., and Das, A. Trimmed maximum likelihood estimation for robust learning in generalized linear models. *arXiv preprint arXiv:2206.04777*, 2022.

Konstantinov, N., Frantar, E., Alistarh, D., and Lampert, C. On the sample complexity of adversarial multi-source pac learning. In *International Conference on Machine Learning*, pp. 5416–5425. PMLR, 2020.

Kothari, P. K., Steinhardt, J., and Steurer, D. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1035–1046, 2018.

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pp. 665–674, Washington, DC, USA, 2016. IEEE Computer Society.

Lecué, G. and Mendelson, S. Performance of empirical risk minimization in linear aggregation. 2016.

Li, J. and Ye, G. Robust gaussian covariance estimation in nearly-matrix multiplication time. *Advances in Neural Information Processing Systems*, 33, 2020.

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *COLT*. arXiv preprint arXiv:1802.07895, 2018.

Liu, L., Shen, Y., Li, T., and Caramanis, C. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

McDonald, J. H. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

Mukhoty, B., Gopakumar, G., Jain, P., and Kar, P. Globally-convergent iteratively reweighted least squares for robust regression problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 313–322, 2019.

Pal, S., Mazumdar, A., Sen, R., and Ghosh, A. On learning mixture of linear regressions in the non-realizable setting. In *International Conference on Machine Learning*, pp. 17202–17220. PMLR, 2022.

Park, Y.-J. and Tuzhilin, A. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 11–18, 2008.

Paschou, P., Lewis, J., Javed, A., and Drineas, P. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 47(12):835–847, 2010.

Pensia, A., Jog, V., and Loh, P.-L. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Philippe, R. 18.s997 high-dimensional statistics. *Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA*, 2015.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Qiao, M. and Valiant, G. Learning discrete distributions from untrusted batches. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pp. 47:1–47:20, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Raghavendra, P. and Yau, M. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 161–180. SIAM, 2020.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. Genetic structure of human populations. *science*, 298 (5602):2381–2385, 2002.

Rousseeuw, P. J. Tutorial to robust statistics. *Journal of chemometrics*, 5(1):1–20, 1991.

Sedghi, H., Janzamin, M., and Anandkumar, A. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 1223–1231, 2016.

Steinhardt, J., Valiant, G., and Charikar, M. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. *Advances in Neural Information Processing Systems*, 29, 2016.

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pp. 2892–2897. PMLR, 2019.

Tukey, J. W. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pp. 448–485, 1960.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G. (eds.), *Compressed Sensing*, pp. 210–268. Cambridge University Press, 2012.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Wax, M. and Ziskind, I. On unique localization of multiple sources by passive sensor arrays. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):996–1000, 1989.

Yi, X., Caramanis, C., and Sanghavi, S. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

Zhong, K., Jain, P., and Dhillon, I. S. Mixed linear regression with multiple components. In *Advances in neural information processing systems (NIPS)*, pp. 2190–2198, 2016.

A. Related Work

Robust Estimation and Regression. Designing estimators which are robust under the presence of outliers has been broadly studied since 1960s (Tukey, 1960; Anscombe, 1960; Huber, 1964). However, most prior works either requires exponential time or have a dimension dependency on the error rate, even for basic problems such as mean estimation. Recently, (Diakonikolas et al., 2019a) proposed a filter-based algorithm for mean estimation which achieves polynomial time and has no dependency on the dimensionality in the estimation error. There has been a flurry of research on robust estimation problems, including mean estimation (Lai et al., 2016; Diakonikolas et al., 2017; Dong et al., 2019; Hopkins et al., 2020a;b; Diakonikolas et al., 2018a), covariance estimation (Cheng et al., 2019; Li & Ye, 2020), linear regression and sparse regression (Bhatia et al., 2015; 2017a; Balakrishnan et al., 2017; Gao, 2020; Prasad et al., 2018; Klivans et al., 2018; Diakonikolas et al., 2019b; Liu et al., 2018; Karmalkar & Price, 2019; Dalalyan & Thompson, 2019; Mukhoty et al., 2019; Diakonikolas et al., 2019c; Karmalkar et al., 2019; Pensia et al., 2020; Cherapanamjeri et al., 2020b), principal component analysis (Kong et al., 2020a; Jambulapati et al., 2020), mixture models (Diakonikolas et al., 2020a; Jia & Vempala, 2019; Kothari et al., 2018; Hopkins & Li, 2018). The results on robust linear regression are particularly related to the setting of this work, though those papers considered non-batch settings and the fraction of good examples $\alpha > 1/2$. (Prasad et al., 2018; Diakonikolas et al., 2019b;c; Pensia et al., 2020; Cherapanamjeri et al., 2020b; Jambulapati et al., 2021) considered the setting when both both covariate x_i and label y_i are corrupted. When there are only label corruptions, (Bhatia et al., 2015; Dalalyan & Thompson, 2019; Kong et al., 2022) achieve nearly optimal rates with $O(d)$ samples. Under the oblivious label corruption model, i.e., the adversary only corrupts a fraction of labels in complete ignorance of the data, (Bhatia et al., 2017b; Suggala et al., 2019) provide a consistent estimator whose approximate error goes to zero as the sample size goes to infinity.

Robust Learning from Batches. (Qiao & Valiant, 2018) introduced the problem of learning discrete distribution from untrusted batches and derived an exponential time algorithm. Subsequent works (Chen et al., 2020b) improved the run-time to quasi-polynomial and (Jain & Orlitsky, 2020a) obtained polynomial time with an optimal sample complexity. (Jain & Orlitsky, 2021; Chen et al., 2020a) extended these results to one-dimensional structured distributions. (Jain & Orlitsky, 2020b; Konstantinov et al., 2020) studied the problem of classification from untrusted batches. (Acharya et al., 2022) studies a closely related problem of learning parameter of Erdős-Rényi random graph when a fraction of nodes are corrupt. All these works focus on different problems than ours and only consider the case when a majority of the data is genuine.

List Decodable Mean Estimation and Regression. List decodable framework was first introduced in (Charikar et al., 2017) to obtain learning guarantees when a majority of data is corrupt. They derived the first polynomial algorithm for list decodable mean estimation under co-variance bound. Subsequent works (Diakonikolas et al., 2020b; Cherapanamjeri et al., 2020c; Diakonikolas et al., 2021a) obtained a better run time. (Diakonikolas et al., 2018b; Kothari et al., 2018) improved the error guarantees, however, under stronger distributional assumptions and has higher sample and time complexities.

(Karmalkar et al., 2019) studies the problem of list-decodable linear regression with batch-size $n = 1$ and derive an algorithm with sample complexity $(d/\alpha)^{O(1/\alpha^4)}$ and runtime $(d/\alpha)^{O(1/\alpha^8)}$. (Raghavendra & Yau, 2020) show a sample complexity of $(d/\alpha)^{O(1/\alpha^4)}$ with runtime $(d/\alpha)^{O(1/\alpha^8)}(1/\alpha)^{\log(1/\alpha)}$. Polynomial time might indeed be impossible for the single sample setting owing to the statistical query lower bounds in (Diakonikolas et al., 2021b).

Mixed Linear Regression. When each batch has only one sample, (i.e. $n = 1$) and contains samples of one of the k regression components the problem becomes the classical mixed linear regression which has been widely studied (Diakonikolas & Kane, 2020; Chen et al., 2020b; Li & Liang, 2018; Sedghi et al., 2016; Zhong et al., 2016; Yi et al., 2016; Chaganty & Liang, 2013). It is worth noting that no algorithm is known to achieve polynomial sample complexity in this setting. The problem is only studied very recently in the batched setting with $n > 1$ by (Kong et al., 2020b;a), where all the samples in the batch are from the same component. (Kong et al., 2020b) proposed a polynomial time algorithm which requires $O(d)$ batches each with size $O(\sqrt{k})$. (Kong et al., 2020a) leveraged sum-of-squares hierarchy to introduce a class of algorithms which is able to trade off the batch size n and the sample complexity. Both of these works assume that the distributions of covariates for all components is identical and Gaussian. Since the above problem is a special case of the list-decodable linear regression, our algorithm is able to recover the k regression components with batch size $n = O(k)$ and $O(d)$ number of batches. Our algorithms allow more general distributions for the covariates than allowed by the Gaussian assumption in the previous works. Further, our algorithms allow the distributions of covariates for the different components to differ. It is worth noting that list-decodable linear regression is a strictly harder problem than mixed linear regression as shown in (Diakonikolas et al., 2021b) and thus our result is incomparable to the ones in the mixed linear regression setting. Leaning mixture of linear dynamical systems has been studied in (Chen & Poor, 2022).

B. Regularity conditions

In this section, we state regularity conditions for genuine data used in proving the guarantees of our algorithm. Before we proceed we will define upper and lower bounds on the clipping parameter κ that are functions of w and other distribution parameters,

Definition B.1. We define the following upper and lower bounds on the clipping parameter κ as a function of w and other distribution parameters:

$$\kappa_{\max} = c_7 C^2 \Big(\sqrt{\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w - w^*)|^2]} + \sigma \Big),$$
$$\kappa_{\min} = \max \Big\{ 8\sqrt{C\, \mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w - w^*)|^2]}, 8\sigma \Big\}.$$

$\kappa_{\max}$ will be used in this section to define our first regularity condition, while $\kappa_{\min}$ will be used in Section C for defining a nice triplet.

Regularity Conditions.

1. For all $\kappa \le \kappa_{\max}$, all unit vectors u and all vectors w

$$\mathbb{E}_G \Big[\big(\nabla f^b(w, \kappa) \cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w, \kappa) \cdot u] \big)^2 \Big] \le c_4 \frac{\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)^2]}{n},$$

2. For all vectors w,

$$\mathbb{E}_G \left[\left(\frac{1}{n} \sum_{i \in [n]} |w \cdot x_i^b - y_i^b| - \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|] \right)^2 \right] \le c_2 \left(\frac{\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]^2}{n} \right).$$

The first regularity condition on the set of good batches G, bounds the mean squared deviation of projections of clipped batch gradients from its true population mean. The regularity condition requires clipping parameter κ to be upper bounded, with the upper bound depending on $\|w - w^*\|$ and σ.

As discussed in Section 5, when $\kappa \to \infty$, the clipping has no effect, and establishing such regularity condition for unclipped gradients would require $\Omega(d^2)$ samples. By using clipping, and ensuring that clipping parameter κ is in the desired range we are able to achieve $\tilde{O}_{n,\alpha}(d)$ sample complexity.

Theorem B.1 characterizes the number of good batches required for regularity condition 1 as a function of the upper bound on κ.

Theorem B.1. *There exist a universal constant c_4 such that for $\mu_{\max} \in [1, \frac{d^4 n^2}{C}]$ and $|G| = \Omega(\mu_{\max}^4 n^2 d \log(d))$, with probability $\ge 1 - \frac{4}{d^2}$, for all unit vectors u, all vectors w and for all $\kappa^2 \le \mu_{\max}(\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)^2])$,*

$$\mathbb{E}_G \Big[\big(\nabla f^b(w, \kappa) \cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w, \kappa) \cdot u] \big)^2 \Big] \le c_4 \frac{\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)^2]}{n}. \tag{12}$$

We prove the above theorem in Section H.

The second regularity condition on the set of good batches G, bounds the mean squared deviation of average absolute error for a batch from its true population mean. Theorem B.2 characterizes the number of good batches required for regularity condition 2.

Theorem B.2. *For $|G| = \Omega(n^2 d \log(d))$ and universal constant $c_2 > 0$, with probability $\ge 1 - \frac{4}{d^2}$, for all vectors w,*

$$\mathbb{E}_G \left[\left(\frac{1}{n} \sum_{i \in [n]} |w \cdot x_i^b - y_i^b| - \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|] \right)^2 \right] \le c_2 \left(\frac{\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|^2]}{n} \right).$$

Proof. Proof of the above theorem is similar to the proof of Theorem B.1, and for brevity, we skip it. □

Combining the two theorems shows that the two regularity conditions hold with high probability with $\tilde{O}_{n,\alpha}(d)$ batches.

Corollary B.3. *For $|G| \geq \Omega_C\big(dn^2 \log(d)\big)$, both regularity conditions hold with probability $\geq 1 - \frac{8}{d^2}$.*

We conclude the sections with the following Lemma which lists some simple consequences of regularity conditions, that we use in later sections.

Lemma B.4. *If regularity conditions hold then*

1. *For all vectors w and for all $\kappa \leq \kappa_{\max}$,*
$$\|\mathrm{Cov}_G(\nabla f^b(w,\kappa))\| \leq c_4 \frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n},$$
2. *For all vectors w*
$$\mathrm{Var}_G\left(\frac{1}{n}\sum_{i\in[n]} |w\cdot x_i^b - y_i^b|\right) \leq c_2\left(\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[|w\cdot x_i^b - y_i^b|]^2}{n}\right).$$
3. *For all $G' \subseteq G$ of size $\geq |G|/2$,*
$$\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)] - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\| \leq \sqrt{2c_4}\frac{\sigma + \sqrt{C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}}{\sqrt{n}}.$$

Proof. The first item in the lemma follows as

$$\begin{aligned}
\|\mathrm{Cov}_G(\nabla f^b(w,\kappa))\| &= \max_{u:\|u\|\leq 1} \mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_G[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] \\
&\leq \max_{u:\|u\|\leq 1} \mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] \\
&\leq c_4\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n},
\end{aligned}$$

where the first inequality follows as the expected squared deviation along the mean is the smallest and the second inequality follows from the first regularity condition.

Similarly, the second item follows from the second regularity condition.

Finally, we prove the last item using the first regularity condition. Let u be any unit vector and $Z^b(u) := \big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2$. Then

$$\|\mathbb{E}_G[Z^b](u)\| = \left\|\frac{1}{|G|}\sum_{b\in G} Z^b(u)\right\| \geq \left\|\frac{1}{|G|}\sum_{b\in B'} Z^b(u)\right\| = \frac{|G'|}{|G|}\|\mathbb{E}_{G'}[Z^b(u)]\| \geq \frac{1}{2}\|\mathbb{E}_{G'}[Z^b(u)]\|,$$

where the first inequality used the fact that $Z^b(u)$ is a positive and the second inequality used $|G'| \geq |G|/2$. Then using the bound on $\|\mathbb{E}_G[Z^b(u)]\|$ in in the first regularity condition, we get

$$\|\mathbb{E}_{G'}[Z^b(u)]\| \leq 2c_4\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n}.$$

Using the Cauchy–Schwarz inequality and the above bound,

$$\mathbb{E}_{G'}[\big|\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big|] = \mathbb{E}_{G'}[\sqrt{Z^b(u)}] \leq \sqrt{\mathbb{E}_{G'}[Z^b(u)]} \leq \sqrt{2c_4\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n}}.$$

Since the above bound holds for each unit vector u, we have

$$\mathbb{E}_{G'}[\big|\nabla f^b(w,\kappa) - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\big|] \leq \sqrt{2c_4\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n}} \leq \sqrt{2c_4}\frac{\sigma + \sqrt{C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}}{\sqrt{n}}.$$

□

C. Guarantees for nice triplet

For completeness, we first restate the conditions a nice triplet (β, κ, w) satisfy.

A triplet (β, κ, w) is *nice* if

(a) β is a nice weight vector, i.e. $\beta^G \geq 3|G|/4$.

(b) $\kappa_{\min} \leq \kappa \leq \kappa_{\max}$.

(c) w is any approximate stationary point w.r.t. β for clipped loss with clipping parameter κ, namely $\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\| \leq \frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}$.

(d) $\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \leq \frac{c_5 C^2 \log^2(2/\alpha)(\sigma^2 + \mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n}$.

In this section, we establish the following guarantees for any nice triplets. In doing so we assume regularity conditions hold for G.

Theorem C.1. *Suppose (β, κ, w) is a nice triplet, $n \geq \max\{32c_4CC_3, \frac{256}{\alpha}c_5C^2{C_3}^2\log^2(2/\alpha)\}$ and regularity conditions holds, then $\|w - w^*\| \leq \mathcal{O}(\frac{C_3C\sigma\log(2/\alpha)}{\sqrt{n\alpha}})$.*

In the remainder of this section, we prove the theorem. First, we provide an overview of the proof and state some auxiliary lemma that we use to prove the theorem.

In this section, we show that for any nice triplet (β, κ, w) if $\|w - w^*\| = \tilde{\Omega}(\sigma/\sqrt{n\alpha})$ then the following lower bound on clipped gradient co-variance, $\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \geq \Omega(\alpha\|w - w^*\|^2)$ holds. For $n = \tilde{\Omega}(\frac{1}{\alpha})$ and $\|w - w^*\| = \tilde{\Omega}(\sigma/\sqrt{n\alpha})$ this lower bound contradicts the upper bound in condition (d). Hence, the theorem concludes that $\|w - w^*\| = \tilde{\mathcal{O}}(\sigma/\sqrt{n\alpha})$.

To show the lower bound $\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \geq \Omega(\alpha\|w - w^*\|^2)$, we first show $\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)]\| = \Omega(\|w - w^*\|) - \tilde{\mathcal{O}}(\sigma/\sqrt{n\alpha})$ in Theorem C.2. Since $\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)]\| = \|\mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\|$, the same bound will hold for the norm of expectation of clipped batch gradients.

When clipping parameter $\kappa \to \infty$ then $\nabla f_i^b(w,\kappa) = \nabla f_i^b(w)$ and for unclipped gradients, a straightforward calculation shows the desired lower bound $\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)]\| = \Omega(\|w - w^*\|)$. However, if κ is too small then clipping may introduce a large bias in the gradients and such a lower bound may no longer hold.

Yet, the lower bound on κ in condition (b) ensures that κ is much larger than the typical error which is of the order $\|w - w^*\| + \sigma$. And when clipping parameter κ is much larger than the typical error, it can be shown that with high probability clipped and unclipped gradients for a random sample from $\mathcal{D}$ would be the same. The next theorem uses this observation and for the case when κ satisfies the lower bound in condition (b) it shows the desired lower bound on the norm of expectation of clipped gradient.

Theorem C.2. *If $\kappa \geq \max\{8\sqrt{C\,\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w - w^*)|^2]}, 8\sigma\}$, then*

$$\left\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)]\right\| \geq \frac{3}{4C_3}\|w - w^*\|.$$

We prove the above theorem in subsection C.1

Since $\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)] = \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]$, the same bound holds for the clipped batch gradients.

Next, in Lemma C.3 we show that for any sufficiently large collection $G' \subseteq G$ of the good batches $\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| \approx \|\mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\|$.

Lemma C.3. *Suppose κ and w are part of a nice triplet, $n \geq 32c_4CC_3$ and regularity conditions holds, then for all $G' \subseteq G$ of size $\geq |G|/2$,*

$$\left\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\right\| \geq \frac{1}{2C_3}\|w - w^*\| - \frac{\sqrt{2c_4}\sigma}{\sqrt{n}}.$$

Proof. From item 3 in Lemma B.4,

$$\begin{aligned}\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)] - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\| &\le \sqrt{2c_4}\cdot\frac{\sigma+\sqrt{C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}}{\sqrt{n}}\\ &\le \frac{\sqrt{2c_4}\sigma}{\sqrt{n}} + \frac{\sqrt{2c_4 C\|w-w^*\|^2\|\Sigma\|}}{\sqrt{n}}\\ &\le \frac{\sqrt{2c_4}\sigma}{\sqrt{n}} + \|w-w^*\|\cdot\frac{\sqrt{2c_4 C}}{\sqrt{n}}.\end{aligned}$$

Using $n \ge 32c_4 C{C_3}^2$,

$$\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)] - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\| \le \frac{1}{4C_3}\|w-w^*\| + \frac{\sqrt{2c_4}\sigma}{\sqrt{n}}.$$

From Theorem C.2, and the observation $\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)] = \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]$, we get

$$\big\|\mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)]\big\| \ge \frac{3}{4C_3}\|w-w^*\|.$$

The lemma follows by combining the above equation using triangle inequality. □

Next, the general bound on the co-variance will be useful in proving Theorem C.1.

Lemma C.4. *For any weight vector β, any set of vectors z^b associated with batches, and any sub-collection of vectors $B' \subseteq \{b \in B : \beta^b \ge 1/2\}$,*

$$Cov_\beta(z^b) \ge \frac{|B'|}{2|B|}\|\mathbb{E}_\beta[z^b] - \mathbb{E}_{B'}[z^b]\|^2.$$

The proof of the lemma appears in Section C.2.

In Theorem C.1 we show that since $\beta^G \ge 3/4|G|$, we can find a sub-collection G' of size $|G|/2$ such that for each $b \in G'$, its weight $\beta^b \ge 1/2$. The we use the previous results for $B' = G'$ and $z = \nabla f^b(w,\kappa)$ to get, $\mathrm{Cov}_\beta(\nabla f^b(w,\kappa)) \ge \frac{|G'|}{4|B|}\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)] - \mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| \ge \frac{|G|}{8|B|}\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)] - \mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| \ge \frac{\alpha}{8}\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)] - \mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\|^2$.

From condition (c) of nice triplets we have $\mathbb{E}_\beta[\nabla f^b(w,\kappa)] \approx 0$ and from Lemma C.3 we have $\mathbb{E}_{G'}[\nabla f^b(w,\kappa)] \gtrsim \|w-w^*\|$. Then from Lemma C.4, we get an upper bound $\mathrm{Cov}_\beta(\nabla f^b(w,\kappa)) \gtrsim \alpha\cdot\|w-w^*\|^2$.

As discussed before, combining this lower bound with the upper bound in condition (d), the theorem concludes $\|w-w^*\| = \tilde{\mathcal{O}}(\sigma/\sqrt{n\alpha})$. Next, we formally prove Theorem C.1 using the above auxiliary lemmas and theorems.

Proof of Theorem C.1. Let $G' := \{b \in G : \beta^b \ge 1/2\}$. Next, we show that $|G'| \ge |G|/2$. To prove it by contradiction assume the contrary that $|G'| < |G|/2$. Then

$$\beta^G = \sum_{b\in G}\beta^b = \sum_{b\in G\setminus G'}\beta^b + \sum_{b\in G'}\beta^b \overset{(a)}{\le} \sum_{b\in G\setminus G'}\frac{1}{2} + \sum_{b\in G'}1 \le \frac{|G\setminus G'|}{2} + |G'| = \frac{|G|-|G'|}{2} + |G'| < 3|G|/4,$$

here (a) follows as the definition of G' implies that for any $b \notin G'$, $\beta^b < 1/2$ and for all batches $\beta^b \le 1$. Above is a contradiction, as we assumed in the Theorem that $\beta^G \ge 3|G|/4$.

Applying Lemma C.4 for $B' = G'$ and $z^b = \nabla f^b(w,\kappa)$ we have

$$\begin{aligned}\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| &\ge \frac{G'}{2|B|}\big(\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| - \|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\|\big)^2\\ &\ge \frac{|G|}{4|B|}\big(\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| - \|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\|\big)^2\\ &\ge \frac{\alpha}{4}\big(\|\mathbb{E}_{G'}[\nabla f^b(w,\kappa)]\| - \|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\|\big)^2. \qquad (13)\end{aligned}$$

In the above equation, using the bound in Lemma C.3 and bound on $\|\mathbb{E}_\beta[\nabla f^b(w,\kappa)]\|$ in condition (c) for nice triplet we get,

$$\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \geq \frac{\alpha}{4}\left(\max\left\{0, \frac{1}{2C_3}\|w-w^*\| - \frac{\sqrt{2c_4}\sigma}{\sqrt{n}} - \frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}\right\}\right)^2.$$

We show that when $\|w-w^*\| \leq \mathcal{O}(\frac{C_3 C\sigma\log(2/\alpha)}{\sqrt{n\alpha}})$, the above upper bound contradicts the following lower bound in condition (d),

$$\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| \leq \frac{c_5 C^2\log^2(2/\alpha)(\sigma^2 + \mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n} \leq \frac{c_5 C^2\log^2(2/\alpha)(\sigma^2+\|w-w^*\|^2)}{n}.$$

To prove the contradiction assume

$$\frac{\|w-w^*\|}{8C_3} > \max\left\{\frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}, \frac{\sqrt{2c_4}\sigma}{\sqrt{n}}, \frac{2\sqrt{c_5}C\sigma\log(2/\alpha)}{\sqrt{n\alpha}}\right\}.$$

Using this lower bound on $\|w-w^*\|$, we lower bound the co-variance. Combining the above lower bound on $\|w-w^*\|$ and equation (13), we get,

$$\begin{aligned}
\|\mathrm{Cov}_\beta(\nabla f^b(w,\kappa))\| &\geq \frac{\alpha}{4}\left(\frac{1}{4C_3}\|w-w^*\|\right)^2 \\
&\geq \frac{\alpha}{4}\left(\frac{2\sqrt{c_5}C\sigma\log(2/\alpha)}{\sqrt{n\alpha}} + \frac{1}{8C_3}\|w-w^*\|\right)^2 \\
&\geq \frac{\alpha}{4}\left(\frac{2\sqrt{c_5}C\sigma\log(2/\alpha)}{\sqrt{n\alpha}}\right)^2 + \frac{\alpha}{4}\left(\frac{1}{8C_3}\|w-w^*\|\right)^2 \\
&\geq \frac{c_5 C^2\log^2(2/\alpha)\sigma^2}{n} + \frac{\alpha}{256}\frac{\|w-w^*\|^2}{{C_3}^2} \\
&\geq \frac{c_5 C^2\log^2(2/\alpha)\sigma^2}{n} + \frac{c_5 C^2\log^2(2/\alpha)\|w-w^*\|^2}{n},
\end{aligned}$$

here the last step used $n \geq \frac{256}{\alpha}c_5 C^2 {C_3}^2\log^2(2/\alpha)$.

This completes the proof of the contradiction. Hence,

$$\frac{\|w-w^*\|}{8C_3} \leq \max\left\{\frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}, \frac{\sqrt{2c_4}\sigma}{\sqrt{n}}, \frac{2\sqrt{c_5}C\sigma\log(2/\alpha)}{\sqrt{n\alpha}}\right\}.$$

The above equation implies $\|w-w^*\| \leq \mathcal{O}(\frac{C_3 C\sigma\log(2/\alpha)}{\sqrt{n\alpha}})$. □

C.1. Proof of Theorem C.2

The following auxiliary lemma will be useful in the proof of the theorem.

Lemma C.5. *For any $z_1 \in \mathbb{R}$, $z_2 > 0$ and a symmetric random variable Z,*

$$\left|\mathbb{E}\left[(z_1+Z) - \frac{(z_1+Z)z_2}{\max(|z_1+Z|, z_2)}\right]\right| \leq 2|z_1|\Pr(Z > z_2 - |z_1|)$$

Proof. We consider $z_1 \geq 0$ and prove the lemma for this case. The proof for $z_1 < 0$ case then follows from the symmetry of the distribution of Z around 0.

The term inside the expectation can be expressed in terms of indicator random variables as follows:

$$\begin{aligned}
&(z_1+Z) - \frac{(z_1+Z)z_2}{\max(|z_1+Z|, z_2)} \\
&= (z_1+Z-z_2)\cdot\mathbb{1}(Z > z_2 - z_1) + (z_1+Z+z_2)\cdot\mathbb{1}(Z < -z_2 - z_1) \\
&= (z_1+Z-z_2)\cdot\mathbb{1}(z_2 - z_1 < Z \leq z_2 + z_1) + (z_1+Z-z_2)\cdot\mathbb{1}(Z > z_2+z_1) + (z_1+Z+z_2)\cdot\mathbb{1}(Z < -z_2-z_1).
\end{aligned}$$

Next, taking the expectation on both sides in the above equation,

$$\begin{aligned}
&\mathbb{E}\left[(z_1+Z)-\frac{(z_1+Z)z_2}{\max(|z_1+Z|,z_2)}\right]\\
&=\mathbb{E}[(z_1+Z-z_2)\cdot\mathbb{1}(z_2-z_1<Z\le z_2+z_1)]+\mathbb{E}[(z_1+Z-z_2)\cdot\mathbb{1}(Z>z_2+z_1)]\\
&\quad+\mathbb{E}[(z_1+Z+z_2)\cdot\mathbb{1}(Z<-z_2-z_1)]\\
&=\mathbb{E}[(z_1+Z-z_2)\cdot\mathbb{1}(z_2-z_1<Z\le z_1+z_2)]+2|z_1|\Pr(Z>z_2+z_1),
\end{aligned}$$

where the last step follows because Z is symmetric and $z_1=|z_1|$ since we assumed $z_1\ge 0$.

Then,

$$\begin{aligned}
\left|\mathbb{E}\left[(z_1+Z)-\frac{(z_1+Z)z_2}{\max(|z_1+Z|,z_2)}\right]\right| &= \mathbb{E}[|z_1+Z-z_2|\cdot\mathbb{1}(z_2-z_1<Z\le z_2+z_1)]+2|z_1|\Pr(Z>z_2+z_1)\\
&\le 2|z_1|\Pr(z_2-z_1<Z\le z_2+z_1)+2|z_1|\Pr(Z>z_2+z_1)\\
&= 2|z_1|\Pr(Z>z_2-z_1).
\end{aligned}$$

□

Next, using the above lemma we prove Theorem C.2.

Proof of Theorem C.2. Consider a random sample (x_i^b,y_i^b) from distribution $\mathcal{D}$. Recall that $n_i^b=y_i^b-w^*\cdot n_i^b$ denote the random noise and is independent of x_i^b.

Consider $(x_i^b\cdot w-y_i^b)x_i^b-\nabla f_i^b(w,\kappa)$. the difference between the unclipped and the clipped gradient for the sample:

$$\begin{aligned}
(x_i^b\cdot w-y_i^b)x_i^b-\nabla f_i^b(w,\kappa) &= (x_i^b\cdot w-y_i^b)x_i^b-\frac{(x_i^b\cdot w-y_i^b)}{|x_i^b\cdot w-y_i^b|\vee\kappa}\kappa x_i^b\\
&=\left((x_i^b\cdot(w-w^*)-n_i^b)x_i^b-\frac{(x_i^b\cdot(w-w^*)-n_i^b)}{|x_i^b\cdot(w-w^*)-n_i^b|\vee\kappa}\kappa\right)x_i^b, \qquad (14)
\end{aligned}$$

where in the last equality we used the relation between x_i^b,y_i^b and n_i^b.

Next, by applying Lemma C.5, we get:

$$\mathbb{E}_{\mathcal{D}}\left[(x_i^b\cdot(w-w^*)-n_i^b)x_i^b-\frac{(x_i^b\cdot(w-w^*)-n_i^b)}{|x_i^b\cdot(w-w^*)-n_i^b|\vee\kappa}\kappa\,\middle|\,x_i^b\right]\le 2|x_i^b\cdot(w-w^*)|\cdot\Pr\big(n_i^b>\kappa-|x_i^b\cdot(w-w^*)|\big),$$

note that in the above expectation x_i^b is fixed and expectation is taken over n_i^b.

Let $Z:=\mathbb{1}\big(|x_i^b\cdot(w-w^*)|\ge\kappa/2\big)$. Observe that $\Pr(n_i^b>\kappa-|(w-w^*)\cdot x_i^b|)\le Z+\Pr(n_i^b>\kappa/2)$. Combining this observation with the above equation, we have:

$$\mathbb{E}_{\mathcal{D}}\left[((w-w^*)\cdot x_i^b-n_i^b)-\frac{((w-w^*)\cdot x_i^b-n_i^b)}{|(w-w^*)\cdot x_i^b-n_i^b|\vee\kappa}\kappa\,\middle|\,x_i^b\right]\le 2|(w-w^*)\cdot x_i^b|\cdot\big(\Pr(n_i^b>\kappa/2)+Z\big). \qquad (15)$$

When $w\ne w^*$ the bound holds trivially. Hence, in the remainder of the proof, we assume $w\ne w^*$. Let $v:=\frac{w-w^*}{\|w-w^*\|}$ and

$Z_i^b := \mathbb{1}\big((|x_i^b \cdot (w - w^*)| \geq \kappa/2) \cup (|n_i^b| \geq \kappa/2)\big)$. Then, for unit vector $v \in \mathbb{R}^d$, we have

$$\begin{aligned}
&|\,\mathbb{E}_{\mathcal{D}}[((w \cdot x_i^b - y_i^b)x_i^b - \nabla f_i^b(w,\kappa)) \cdot v]| \\
&= |\,\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\mathcal{D}}[((w \cdot x_i^b - y_i^b)x_i^b - \nabla f_i^b(w,\kappa)) \cdot v | x_i^b]]| \\
&\leq \mathbb{E}_{\mathcal{D}}[|\,\mathbb{E}_{\mathcal{D}}[((w \cdot x_i^b - y_i^b)x_i^b - \nabla f_i^b(w,\kappa)) \cdot v \big| x_i^b]]| \\
&\overset{(a)}{\leq} \mathbb{E}_{\mathcal{D}}\big[2|(w - w^*) \cdot x_i^b| \cdot |x_i^b \cdot v|\big(Z + \Pr(n_i^b > \kappa/2)\big)\big] \\
&\overset{(b)}{\leq} \mathbb{E}_{\mathcal{D}}\left[\frac{2|(w - w^*) \cdot x_i^b|^2}{\|w - w^*\|}\big(Z + \Pr(n_i^b > \kappa/2)\big)\right] \\
&\leq \frac{2}{\|w - w^*\|}\big(\mathbb{E}_{\mathcal{D}}[Z \cdot |(w - w^*) \cdot x_i^b|^2] + \Pr(n_i^b > \kappa/2)\,\mathbb{E}_{\mathcal{D}}[|(w - w^*) \cdot x_i^b|^2]\big),
\end{aligned} \tag{16}$$

here (a) follows from Equation (14) and Equation (15), and (b) follows from the definition of vector v. Next, we bound the two terms on the right one by one. We start with the first term:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[Z \cdot |(w - w^*) \cdot x_i^b|^2] &\overset{(a)}{\leq} \big(\mathbb{E}[(Z)^2] \cdot \mathbb{E}_{\mathcal{D}}[(x_i^b \cdot (w - w^*))^4]\big)^{1/2} \\
&\overset{(b)}{\leq} \big(\mathbb{E}[Z] \cdot C\,\mathbb{E}_{\mathcal{D}}[(x_i^b \cdot (w - w^*))^2]^2\big)^{1/2} \\
&\overset{(c)}{\leq} \big(C \Pr[|x_i^b \cdot (w - w^*)| \geq \kappa/2]\big)^{1/2} \cdot \mathbb{E}_{\mathcal{D}}[(x_i^b \cdot (w - w^*))^2],
\end{aligned} \tag{17}$$

where (a) used the Cauchy-Schwarz inequality, (b) used the fact that Z is an indicator random variable, hence, $Z^2 = Z$ and $L4 - L2$ hypercontractivity, and (c) follows from the definition of Z.

Applying the Markov inequality to $(n_i^b)^2$ we get:

$$\Pr[|n_i^b| \geq \kappa/2] \leq \frac{\mathbb{E}_{\mathcal{D}}[(n_i^b)^2]}{(\kappa/2)^2} \leq \frac{\sigma^2}{(\kappa/2)^2}. \tag{18}$$

Similarly, applying the Markov inequality to $|x_i^b \cdot (w - w^*)|^4$ yields:

$$\Pr[|x_i^b \cdot (w - w^*)| \geq \kappa/2] \leq \frac{\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w - w^*)|^4]}{(\kappa/2)^4} \leq \frac{C\,\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w - w^*)|^2]^2}{(\kappa/2)^4}, \tag{19}$$

where the last inequality uses $L4 - L2$ hypercontractivity.

Combining Equations (16), (17), (18) and (19), we have

$$|\,\mathbb{E}_{\mathcal{D}}[((w \cdot x_i^b - y_i^b)x_i^b - \nabla f_i^b(w,\kappa)) \cdot v]| \leq \frac{8\,\mathbb{E}_{\mathcal{D}}[(x_i^b \cdot (w - w^*))^2]}{\kappa^2 \|w - w^*\|}\big(C\,\mathbb{E}_{\mathcal{D}}[(x_i^b \cdot (w - w^*))^2] + \sigma^2\big).$$

Next,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[(w \cdot x_i^b - y_i^b)x_i^b \cdot v] &\overset{(a)}{=} \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b - n_i^b)x_i^b \cdot v] \\
&\overset{(b)}{=} \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)x_i^b \cdot v] - \mathbb{E}_{\mathcal{D}}[n_i^b] \cdot \mathbb{E}_{\mathcal{D}}[x_i^b \cdot v] \\
&\overset{(c)}{=} \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)x_i^b \cdot v] \\
&\overset{(d)}{=} \frac{\mathbb{E}_{\mathcal{D}}\big[((w - w^*) \cdot x_i^b)^2\big]}{\|w - w^*\|},
\end{aligned}$$

here (a) follows from the relationship between x_i^b, y_i^b and n_i^b, (b) follows from as x_i^b and n_i^b are independent, (c) uses $\mathbb{E}_{\mathcal{D}}[n_i^b] = 0$ and (d) follows from the definition of v.

Combining the previous two equations using the triangle inequality:

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)\cdot v]| &\geq |\mathbb{E}_{\mathcal{D}}[(w\cdot x_i^b - y_i^b)x_i^b\cdot v]| - |\mathbb{E}_{\mathcal{D}}[((w\cdot x_i^b - y_i^b)x_i^b - \nabla f_i^b(w,\kappa))\cdot v]| \\
&\geq \mathbb{E}_{\mathcal{D}}\left[\frac{((w-w^*)\cdot x_i^b)^2}{\|w-w^*\|}\right]|\left(1-\frac{8}{\kappa^2}\left(C\,\mathbb{E}_{\mathcal{D}}[(x_i^b\cdot(w-w^*))^2]+\sigma^2\right)\right) \\
&\geq \mathbb{E}_{\mathcal{D}}\left[\frac{((w-w^*)\cdot x_i^b)^2}{\|w-w^*\|}\right]\left(1-\frac{1}{4}\right) \\
&\geq \frac{3}{4}\|w-w^*\|\cdot\frac{\|\Sigma\|}{C_3} = \frac{3}{4C_3}\|w-w^*\|,
\end{aligned}$$

here the second last inequality follows from lower bound on κ.

The theorem then follows by observing,

$$\|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)]\| \geq \max_{\|u\|=1}|\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)\cdot u]| \geq |\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)\cdot v]| \geq \frac{3}{4C_3}\|w-w^*\|.$$

□

C.2. Proof of Lemma C.4

Proof. Note that

$$\begin{aligned}
\|\mathrm{Cov}_\beta(z^b)\| &= \left\|\sum_{b\in B}\frac{\beta^b}{\beta^B}(z^b-\mathbb{E}_\beta[z^b])(z^b-\mathbb{E}_\beta[z^b])^\intercal\right\| \\
&\geq \left\|\sum_{b\in B'}\frac{\beta^b}{\beta^B}(z^b-\mathbb{E}_\beta[z^b])(z^b-\mathbb{E}_\beta[z^b])^\intercal\right\| \\
&\overset{(a)}{\geq} \left\|\sum_{b\in B'}\frac{1}{2|B|}(z^b-\mathbb{E}_\beta[z^b])(z^b-\mathbb{E}_\beta[z^b])^\intercal\right\| \\
&\overset{(b)}{\geq} \frac{1}{2|B|}\left\||B'|(\mathbb{E}_{B'}[z^b]-\mathbb{E}_\beta[z^b])(\mathbb{E}_{B'}[]z^b]-\mathbb{E}_\beta[z^b])^\intercal\right\| \\
&= \frac{|B'|}{2|B|}\|\mathbb{E}_\beta[z^b]-\mathbb{E}_{B'}[z^b]\|^2,
\end{aligned}$$

where (a) used $\beta^b \geq 1/2$ for $b\in B'$ and the trivial bound $\beta^B \leq |B|$ and (b) follows from the fact that any Z,

$$\left\|\sum_{b\in B'}(z^b-Z)(z^b-Z)^\intercal\right\| \geq |B'|\cdot\left\|\left(\mathbb{E}_{B'}[z^b]-Z\right)\left(\mathbb{E}_{B'}[z^b]-Z\right)^\intercal\right\|.$$

We complete the proof of the lemma by proving the above fact.

$$\begin{aligned}
&\left\|\sum_{b\in B'}(z^b-Z)(z^b-Z)^\intercal\right\| = \left\|\sum_{b\in B'}(z^b-\mathbb{E}_{B'}[z^b]+\mathbb{E}_{B'}[z^b]-Z)(z^b-\mathbb{E}_{B'}[z^b]+\mathbb{E}_{B'}[z^b]-Z)^\intercal\right\| \\
&\overset{(a)}{=} \left\|\sum_{b\in B'}\left((z^b-\mathbb{E}_{B'}[z^b])(z^b-\mathbb{E}_{B'}[z^b])^\intercal+(\mathbb{E}_{B'}[z^b]-Z)(\mathbb{E}_{B'}[z^b]-Z)^\intercal\right)\right\| \\
&\overset{(b)}{\geq} |B'|\cdot\left\|(\mathbb{E}_{B'}[z^b]-Z)(\mathbb{E}_{B'}[z^b]-Z)^\intercal\right\|,
\end{aligned}$$

here (a) follows as $\sum_{b\in B'}z^b = |B'|\,\mathbb{E}_{B'}[z^b]$ and hence, $\sum_{b\in B'}(z^b-\mathbb{E}_{B'}[z^b])(\mathbb{E}_{B'}[z^b]-Z)^\intercal = \sum_{b\in B'}(\mathbb{E}_{B'}[z^b]-Z)(z^b-\mathbb{E}_{B'}[z^b])^\intercal = 0$, and (b) follows as $(z^b-\mathbb{E}_{B'}[z^b])(z^b-\mathbb{E}_{B'}[z^b])^\intercal$ are positive semi-definite matrices.

□

Algorithm 2 FINDCLIPPINGPPARAMETER

Input: Set B, β, σ, $a_1 \geq 1$, a_2 data $\{\{(x_i^b, y_i^b\}_{i\in[n]}\}_{b\in B}$.
$\kappa \leftarrow \infty$
while True **do**
  $w_\kappa \leftarrow$ any approximate stationary point of clipped losses $\{f^b(\cdot\,, \kappa)\}$ w.r.t. weight vector β such that $\|\,\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]\| \leq \frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}$
  $\kappa_{new} \leftarrow \max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\}$.
  if $\kappa_{new} \geq \kappa/2$ **then**
    Break
  end if
  $\kappa \leftarrow \kappa_{new}$
end while
Return(κ, w_κ)

D. Subroutine FINDCLIPPINGPARAMETER and its analysis

Theorem D.1. *For any weight vector β, $a_1 \geq 1$, and $a_2 > 0$, Algorithm* FINDCLIPPINGPARAMETER *runs at most* $\log\Big(\mathcal{O}\Big(\frac{\max_{i,b}|y_i^b|}{\sigma}\Big)\Big)$ *iterations of the while loop and returns κ and w_κ such that*

1. *w_κ is a (approximate) stationary point for $\{f^b(\cdot, \kappa)\}$ w.r.t. weight vector β such that*

$$\|\,\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]\| \leq \frac{\log(2/\alpha)\sigma}{8\sqrt{n\alpha}}.$$

2. $\max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\} \leq \kappa \leq 2\max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\}$.

3. $\max\Big\{\frac{a_1}{2}\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big], a_2\sigma\Big\} \leq \kappa \leq \max\Big\{4a_1^2\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big], a_2\sigma\Big\}$.

Proof. First, we bound the number of iterations of the while loop. Since w_κ is a stationary point for $f^b(.,\kappa)$, hence its will achieve a smaller loss than $w = 0$, hence $\mathbb{E}_\beta[f^b(w_\kappa, \kappa)] \leq \mathbb{E}_\beta[f^b(0, \kappa)]$. And, since the clipped loss is smaller than unclipped loss, $\mathbb{E}_\beta[f^b(0, \kappa)] \leq \mathbb{E}_\beta[f^b(0)] = \mathbb{E}_\beta[\frac{1}{n}\sum_{i\in[n]}(y_i^b)^2] \leq \max_{i,b}(y_i^b)^2$. Therefore after the first iteration $\kappa \leq \max\{a_1 \max_{i,b}|y_i^b|, a_2\sigma\}$. Also in each iteration apart from the last one κ decreases by a factor 2 and κ can't be smaller than $a_2\sigma$. Hence, the number of iterations between the first one and the last one are at most $\log(\frac{a_1 \max_{i,b}|y_i^b|}{a_2\sigma}))$. Therefore the total number of iterations are at most $\log(\frac{a_1 \max_{i,b}|y_i^b|}{a_2\sigma})) + 2$.

The first item follows from the definition of w_κ in the subroutine FINDCLIPPINGPPARAMETER.

Next to prove the lower bound in item 2 we prove the claim that if in an iteration $\kappa \geq \max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\}$ then the same condition will hold in the next iteration.

The condition $\kappa \geq \max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\}$ in the claim implies that $\kappa \geq \kappa_{new}$. Then from the definition of clipped loss, for each w and each b we have $f^b(w, \kappa) \geq f^b(w, \kappa_{new})$. It follows that $\mathbb{E}_\beta[f^b(w_\kappa, \kappa)] \geq \mathbb{E}_\beta[f^b(w_\kappa, \kappa_{new})]$. And further $w_{\kappa_{new}}$ is stationary point for $f^b(.,\kappa_{new})$, hence it will achieve a smaller loss, $\mathbb{E}_\beta[f^b(w_{\kappa_{new}}, \kappa_{new})] \leq \mathbb{E}_\beta[f^b(w_\kappa, \kappa_{new})]$. Therefore, $\mathbb{E}_\beta[f^b(w_{\kappa_{new}}, \kappa_{new})] \leq \mathbb{E}_\beta[f^b(w_\kappa, \kappa)]$. Hence, $\kappa_{new} = \max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]}, a_2\sigma\Big\} \geq \max\Big\{a_1\sqrt{\mathbb{E}_\beta[f^b(w_{\kappa_{new}}, \kappa_{new})]}, a_2\sigma\Big\}$. This completes the proof of the claim.

Since the initial value of κ is infinite the claim must hold in the first iteration, and therefore in each iteration thereafter. Therefore it must hold in the iteration when the algorithm terminates. This completes the proof of the lower bound in item 2.

The upper bound in the second item follows by observing that when the algorithm ends $\kappa \leq 2\kappa_{new}$ and $\kappa_{new} = a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa, \kappa)]} + a_2\sigma$.

Finally, we prove item 3 using item 2. We start by proving the lower bound in item 3. From the lower bound in item 2, we have, $\kappa \geq a_2\sigma$. Then to complete the proof of the lower bound in item 3, it suffices to prove $\kappa > \frac{a_1}{2}\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big]$. To prove this by contradiction suppose $\kappa < \frac{a_1}{2}\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big]$. Then

$$
\begin{aligned}
&\mathbb{E}_\beta[f^b(w_\kappa,\kappa)]\\
&= \mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]} f_i^b(w_\kappa,\kappa)\right]\\
&= \frac{1}{n}\sum_{i\in[n]}\mathbb{E}_\beta\left[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot\frac{(w_\kappa \cdot x_i^b - y_i^b)^2}{2} + \mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot\left(\kappa|w_\kappa \cdot x_i^b - y_i^b| - \frac{\kappa^2}{2}\right)\right]\\
&\geq \frac{1}{n}\sum_{i\in[n]}\left(\mathbb{E}_\beta\left[\left(\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot\frac{(w_\kappa \cdot x_i^b - y_i^b)^2}{2}\right] + \mathbb{E}_\beta\left[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot\left(\frac{\kappa|w_\kappa \cdot x_i^b - y_i^b|}{2}\right)\right)\right]\right)\\
&\overset{(a)}{\geq} \frac{1}{2n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]^2 + \frac{\kappa}{2n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\\
&\overset{(b)}{\geq} \frac{1}{2}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\right)^2 + \frac{\kappa}{2n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\\
&\overset{(c)}{\geq} \frac{\kappa}{a_1}\left(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\right) + \frac{\kappa}{2n}\sum_{i\in[n]}\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\\
&\overset{(d)}{\geq} \frac{\kappa}{2a_1 n}\sum_{i\in[n]}\big(\mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| \leq \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big] + \mathbb{E}_\beta\big[\mathbb{1}(|w_\kappa \cdot x_i^b - y_i^b| > \kappa)\cdot|w_\kappa \cdot x_i^b - y_i^b|\big]\big)\\
&= \frac{\kappa}{2a_1}\,\mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\right]\\
&\overset{(e)}{\geq} \frac{\kappa^2}{a_1^2},
\end{aligned}
\tag{20}
$$

here (a) and (b) follows the Cauchy-Schwarz inequality, (c) and (e) follows from our assumption $\kappa < \frac{a_1}{2}\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big]$ and (d) follows since $a_1 \geq 1$.

This contradicts the lower bound $\kappa \geq a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa,\kappa)]}$ in item 2. Hence we conclude, $\kappa \geq \frac{a_1}{2}\,\mathbb{E}_\beta\Big[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\Big]$. This completes the proof of the lower bound in item 3.

Next, we prove the upper bound in item 3. We consider two cases. For the case when $a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa,\kappa)]} \leq a_2\sigma$ then upper bound in item 3 follows from the upper bound in item 2. Next we prove for the other case, when $a_1\sqrt{\mathbb{E}_\beta[f^b(w_\kappa,\kappa)]} > a_2\sigma$. For this case item 2 implies $\mathbb{E}_\beta[f^b(w_\kappa,\kappa)] \geq \frac{\kappa^2}{4a_1^2}$.

Next, from the definition of $f^b(w,\kappa)$,

$$
\mathbb{E}_\beta[f^b(w_\kappa,\kappa)] = \mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]} f_i^b(w_\kappa,\kappa)\right] \leq \mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]}\kappa|w_\kappa \cdot x_i^b - y_i^b|\right] \leq \kappa\,\mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\right]. \tag{21}
$$

Combining the above equation and $\mathbb{E}_\beta[f^b(w_\kappa,\kappa)] \geq \frac{\kappa^2}{4a_1^2}$, we get,

$$
\frac{\kappa^2}{4a_1^2} \leq \kappa\,\mathbb{E}_\beta\left[\frac{1}{n}\sum_{i\in[n]}|w_\kappa \cdot x_i^b - y_i^b|\right].
$$

The upper bound in item 3 then follows from the above equation. □

E. Correctness of estimated parameters for nice weight vectors

For batch $b \in B$, let $v^b(w) := \frac{1}{n}\sum_{i\in[n]} |w \cdot x_i^b - y_i^b|$. Since w will be fixed in the proofs, we will often denote $v^b(w)$ as v^b.

In this section, we state and prove Theorems E.1, E.2 and E.4. For any triplet with a nice weight vector, Theorem E.1 ensures the correctness of parameters calculated for Type-1 use of MULTIFILTER. For any triplet with a nice weight vector, Theorem E.4 ensures the correctness of parameters calculated for the case when it gets added to M or goes through Type-2 use of MULTIFILTER. Theorem E.2 serves as an intermediate step in proving Theorem E.4.

Theorem E.1. *In Algorithm 1 if the weight vector β is such that $\beta^G \geq 3|G|/4$, $n \geq (16)^2 c_2 C$, and Theorem B.2's conclusion holds, then for any w, the parameter θ_1 computed in the subroutine satisfies*

$$\theta_1 \geq c_2\left(\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]^2}{n}\right),$$

where c_2 is the same universal positive constant as item 2 in Lemma B.4.

Proof. To prove the theorem we first show that θ_0 calculated in the algorithm is $\geq \frac{7\,\mathbb{E}_{\mathcal{D}}[|w\cdot x_i^b - y_i^b|]}{8} - \frac{\sigma}{8\sqrt{C}}$.

Let MED denote median of the set $\{v^b : b \in G\}$. From Theorem B.2 and Markov's inequality, it follows that

$$\begin{aligned}\left|\text{MED} - \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]\right| &\leq 2\sqrt{c_2\left(\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]^2}{n}\right)} \\ &\leq \frac{\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{8} + \frac{\sigma}{8\sqrt{C}}. \end{aligned} \tag{22}$$

where the last inequality uses $n \geq (16)^2 c_2 C$. It follows that

$$\text{MED} \geq \frac{7\,\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{8} - \frac{\sigma}{8\sqrt{C}}.$$

Then to complete the proof we show that MED $\leq \theta_0$. Note that

$$\sum_{b\in G: v^b < \text{MED}} \beta^b \leq |\{b \in G : v^b < \text{MED}\}| < \frac{|G|}{2}.$$

Then,

$$\sum_{b\in B: v^b \geq \text{MED}} \beta^b \geq \sum_{b\in G: v^b \geq \text{MED}} \beta^b = \sum_{b\in G} \beta^b - \sum_{b\in G: v^b < \text{MED}} \beta^b > \beta^G - \frac{|G|}{2} \geq \frac{3|G|}{4} - \frac{|G|}{2} \geq \frac{|G|}{4}. \tag{23}$$

And since from the definition of θ_0, we have $\sum_{b:v^b>\theta_0} \beta^b \leq \alpha|B|/4 \leq \frac{|G|}{4}$, it follows that MED $\leq \theta_0$.

Therefore, $\theta_0 \geq \frac{7\,\mathbb{E}_{\mathcal{D}}[|w\cdot x_i^b - y_i^b|]}{8} - \frac{\sigma}{8\sqrt{C}}$. The lower bound in the theorem on θ_1 then follows from the relation between θ_0 and θ_1. □

Theorem E.2. *Suppose regularity conditions holds, and β, w and n satisfy $n \geq \max\{\frac{(32)^2 c_3 c_2 C \log^2(2/\alpha)}{\alpha}, (16)^2 c_2 C\}$, $\beta^G \geq 3|G|/4$, and*

$$\mathit{Var}_\beta\big(v^b(w)\big) \leq c_3 \log^2(2/\alpha)\theta_1,$$

then

$$\frac{3\,\mathbb{E}_{\mathcal{D}}[|(w - w^*)\cdot x_i^b|]}{4} - \sigma \leq \mathbb{E}_\beta\big[v^b(w)\big] \leq \frac{4\,\mathbb{E}_{\mathcal{D}}[|(w - w^*)\cdot x_i^b|]}{3} + 2\sigma.$$

In proving Theorem E.2 the following auxiliary lemma will be useful. We prove this lemma in Subsection E.1.

Lemma E.3. *Let Z be any random variable over the reals. For any $z \in \mathbb{R}$, such that $\Pr[Z > z] \le 1/2$, we have*

$$z - \sqrt{\frac{\mathrm{Var}(Z)}{\Pr[Z \ge z]}} \le \mathbb{E}[Z] \le z + \sqrt{2\,\mathrm{Var}(Z)}.$$

and for all $z \in Z$,

$$|\,\mathbb{E}[Z] - z| \le \sqrt{\frac{\mathrm{Var}(Z)}{\min\{\Pr[Z \le z], \Pr[Z \ge z], 0.5\}}}.$$

Now we prove Theorem E.2 using the above Lemma.

Proof of Theorem E.2. Let MED denote median of the set $\{v^b : b \in G\}$. In Equation (23) we showed,

$$\sum_{b \in B: v^b \ge \mathrm{MED}} \beta^b \ge \frac{|G|}{4}.$$

Hence,

$$\frac{\sum_{b \in B: v^b \ge \mathrm{MED}} \beta^b}{\beta^B} \ge \frac{|G|}{4|B|} \ge \frac{\alpha}{4}.$$

Similarly, by symmetry, one can show

$$\frac{\sum_{b \in B: v^b \le \mathrm{MED}} \beta^b}{\beta^B} \ge \frac{\alpha}{4}.$$

Then from the second bound in Lemma E.3,

$$|\,\mathbb{E}_\beta[v^b] - \mathrm{MED}| \le \sqrt{\frac{4\,\mathrm{Var}_\beta[v^b]}{\alpha}}. \tag{24}$$

From Equation (22), the above equation, and the triangle inequality,

$$|\,\mathbb{E}_\beta[v^b] - \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]| \le \sqrt{\frac{4\,\mathrm{Var}_\beta[v^b]}{\alpha}} + \frac{\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{8} + \frac{\sigma}{8\sqrt{C}}. \tag{25}$$

Next, from the definition of θ_0, we have $\sum_{b: v^b \ge \theta_0} \beta^b \ge \alpha|B|/4$ and $\sum_{b: v^b > \theta_0} \beta^b < \alpha|B|/4$. Then

$$\frac{\sum_{b: v^b \ge \theta_0} \beta^b}{\beta^B} \ge \frac{\alpha|B|}{4\beta^B} \ge \frac{\alpha|B|}{4|B|} \ge \frac{\alpha}{4},$$

and

$$\frac{\sum_{b: v^b > \theta_0} \beta^b}{\beta^B} < \frac{\alpha|B|}{4\beta^B} \le \frac{\alpha|B|}{4\beta^G} \le \frac{\alpha|B|}{4(3|G|/4)} \le \frac{1}{3}.$$

Then from the first bound in Lemma E.3,

$$\theta_0 - \sqrt{\frac{4\,\mathrm{Var}_\beta[v^b]}{\alpha}} \le \mathbb{E}_\beta[v^b]. \tag{26}$$

In this lemma, we had assumed the following bound on the variance of v^b,

$$\mathrm{Var}_\beta[v^b] \le c_3 \log^2(2/\alpha)\theta_1.$$

Next,

$$\frac{\mathrm{Var}_\beta[v^b]}{\alpha} \le \frac{c_3 \log^2(2/\alpha)\theta_1}{\alpha} = \frac{c_3 \log^2(2/\alpha) c_2(\sigma^2 + (2\sqrt{C}\theta_0 + \sigma)^2)}{n\alpha} \le \frac{(\sigma^2 + 4C\theta_0^2 + 2\sigma^2)}{32^2 C} \le \frac{\sigma^2}{256C} + \frac{\theta_0^2}{256},$$

here the equality follows from the relation between θ_0 and θ_1 and the first inequality follows as $n \ge \frac{(32)^2 C c_3 c_2 \log^2(2/\alpha)}{\alpha}$.

Then

$$\sqrt{\frac{\mathrm{Var}_\beta[v^b]}{\alpha}} \le \sqrt{\frac{\sigma^2}{256C} + \frac{\theta_0^2}{256}} \le \frac{\sigma}{16\sqrt{C}} + \frac{\theta_0}{16} \le \frac{\sigma}{16\sqrt{C}} + \frac{1}{16}\left(\mathbb{E}_\beta[v^b] + 2\sqrt{\frac{\mathrm{Var}_\beta[v^b]}{\alpha}}\right),$$

here the second inequality used $\sqrt{a^2+b^2} \le |a| + |b|$ and the last inequality used (26). From the above equation, it follows that

$$\sqrt{\frac{\mathrm{Var}_\beta[v^b]}{\alpha}} \le \frac{\sigma}{14\sqrt{C}} + \frac{1}{14}\mathbb{E}_\beta[v^b].$$

Combining the above bound and Equation (25)

$$|\mathbb{E}_\beta[v^b] - \mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]| \le \frac{\sigma}{7\sqrt{C}} + \frac{1}{7}\mathbb{E}_\beta[v^b] + \frac{\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{8} + \frac{\sigma}{8\sqrt{C}}.$$

From the above equation it follows that

$$\frac{49\,\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{64} - \frac{15\sigma}{64\sqrt{C}} \le \mathbb{E}_\beta[v^b] \le \frac{21\,\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]}{16} + \frac{5\sigma}{16\sqrt{C}}. \tag{27}$$

Finally, we upper bound and lower bound $\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|]$ to complete the proof. To prove the upper bound, note that,

$$\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|] = \mathbb{E}_{\mathcal{D}}[|(w-w^*) \cdot x_i^b - n_i^b|] \le \mathbb{E}_{\mathcal{D}}[|(w-w^*) \cdot x_i^b|] + \mathbb{E}_{\mathcal{D}}[|n_i^b|] \le \mathbb{E}_{\mathcal{D}}[|(w-w^*) \cdot x_i^b|] + \sigma,$$

here the last inequality used $\mathbb{E}_{\mathcal{D}}[|n_i^b|] \le \sqrt{\mathbb{E}_{\mathcal{D}}[|n_i^b|^2]}$. Combining the above upper bound with the upper bound in (27) and using $C \ge 1$ proves the upper bound in the lemma. Similarly, we can show

$$\mathbb{E}_{\mathcal{D}}[|w \cdot x_i^b - y_i^b|] \ge \mathbb{E}_{\mathcal{D}}[|(w-w^*) \cdot x_i^b|] - \sigma,$$

Combining the above lower bound with the lower bounds in (27) and using $C \ge 1$ proves the lower bound in the lemma. □

Theorem E.4. *Suppose regularity conditions holds, and β, w and n satisfy $n \ge \max\{\frac{(32)^2 c_3 c_2 C \log^2(2/\alpha)}{\alpha}, (16)^2 c_2 C\}$, $\beta^G \ge 3|G|/4$, and*

$$\mathit{Var}_\beta\left(\frac{1}{n}\sum_{i\in[n]} |w \cdot x_i^b - y_i^b|\right) \le c_3 \log^2(2/\alpha)\theta_1,$$

then for κ, w returned by subroutine FINDCLIPPINGPARAMETER *and θ_2 calculated by* MAINALGORITHM*, we have*

1. $c_4 \frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n} \le \theta_2 \le \frac{c_6 C^2(\sigma^2 + \mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2)}{n}$*, where c_4 is the same positive constant as in item 1 of Lemma B.4 and c_6 is some other positive universal constant.*
2. $\max\{8\sqrt{C\,\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w-w^*)|^2]}, 8\sigma\} \le \kappa$ *and* $\kappa \le c_7 C^2\left(\sqrt{\mathbb{E}_{\mathcal{D}}[|x_i^b \cdot (w-w^*)|^2]} + \sigma\right)$*, where c_7 is some other positive universal constant.*

Note that the range of κ in item 2 of the above Theorem is the same as that in (b).

In proving the theorem the following lemma will be useful.

Lemma E.5. *For any vectors u, we have*

$$\sqrt{\frac{\mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|^2]}{8C}} \le \mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|] \le \sqrt{\mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|^2]}$$

We prove the above auxiliary lemma in Section E.2 using the Cauchy-Schwarz inequality for the upper bound and $L4 - L2$ hypercontractivity for the lower bound.

Next, we prove Theorem E.4 using the above lemma and Theorem E.2.

Proof of Theorem E.4. We start by proving the first item. For convenience, we recall the definition of θ_2 in (8),

$$\theta_2 = \tfrac{c_4}{n}\Big(\sigma^2 + 16C^2\big(\mathbb{E}_{\beta}[v^b] + \sigma\big)^2\Big).$$

The upper bound in the item follows from this definition of θ_2 and the upper bound on $\mathbb{E}_{\beta}[v^b]$ in Lemma E.2.

Using the lower bound bound on $\mathbb{E}_{\beta}[v^b]$ in Lemma E.2 and definition of θ_2,

$$\theta_2 \ge \tfrac{c_4}{n}\big(\sigma^2 + 9C^2\,\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]^2\big) \ge \tfrac{c_4}{n}\big(\sigma^2 + \tfrac{9}{8}C^2\,\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|^2]\big),$$

where the last step used Lemma E.5. This completes the proof of lower bound item 1.

Next, we prove item 2. From Theorem D.1,

$$\max\Big\{\tfrac{a_1}{2}\,\mathbb{E}_{\beta}\Big[\tfrac{1}{n}\textstyle\sum_{i\in[n]}|w\cdot x_i^b - y_i^b|\Big], a_2\sigma\Big\} \le \kappa \le \max\Big\{4a_1^2\,\mathbb{E}_{\beta}\Big[\tfrac{1}{n}\textstyle\sum_{i\in[n]}|w\cdot x_i^b - y_i^b|\Big], a_2\sigma\Big\}.$$

Since for any $a, b > 0$, $(a+b)/2 \le \max(a,b) \le a+b$. Then from the above bound,

$$\tfrac{a_1}{4}\,\mathbb{E}_{\beta}\Big[\tfrac{1}{n}\textstyle\sum_{i\in[n]}|w_\kappa\cdot x_i^b - y_i^b|\Big] + \tfrac{a_2\sigma}{2} \le \kappa \le 4a_1^2\,\mathbb{E}_{\beta}\Big[\tfrac{1}{n}\textstyle\sum_{i\in[n]}|w_\kappa\cdot x_i^b - y_i^b|\Big] + a_2\sigma.$$

Using the bound on $\mathbb{E}_{\beta}[v^b]$ in Lemma E.2 in the above equation

$$\tfrac{a_1}{4}\Big(\tfrac{3\,\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]}{4} - \tfrac{\sigma}{2}\Big) + \tfrac{a_2\sigma}{2} \le \kappa \le 4a_1^2\Big(\tfrac{4\,\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|]}{3} + 2\sigma\Big) + a_2\sigma.$$

Using Lemma E.5, and the above equation,

$$\tfrac{3a_1}{32\sqrt{2C}}\sqrt{\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|^2]} + \tfrac{(4a_2-a_1)\sigma}{8} \le \kappa \le \tfrac{16a_1^2}{3}\sqrt{\mathbb{E}_{\mathcal{D}}[|(w-w^*)\cdot x_i^b|^2]} + (8a_1^2 + a_2)\sigma.$$

The upper bound and lower bound in item 2 then follow by using the values $a_1 = \frac{256C\sqrt{2}}{3}$ and $a_2 = \frac{a_1}{4} + 64$. □

E.1. Proof of Lemma E.3

Proof of Lemma E.3. We only prove the first statement as the second statement and then follow from the symmetry.

We start by proving the upper bound in the first statement. We consider two cases, $\mathbb{E}[Z] \le z$ and $\mathbb{E}[Z] > z$. For the first case, the upper bound automatically follows. Next, we prove the second case. In this case,

$$\mathrm{Var}(Z) = \mathbb{E}[(Z-\mathbb{E}[Z])^2] \ge \mathbb{E}[\mathbb{1}(Z\le z)(Z-\mathbb{E}[Z])^2] \ge \mathbb{E}[\mathbb{1}(Z\le z)(z-\mathbb{E}[Z])^2] = \Pr[Z\le z](z-\mathbb{E}[Z])^2.$$

Then using $\Pr[Z \le z] = 1 - \Pr[Z > z] \ge 1/2$, we get

$$\mathrm{Var}(Z) \ge \frac{(z-\mathbb{E}[Z])^2}{2}.$$

The upper bound from the above equation.

Next, we prove the lower bound. Again, we consider two cases, $\mathbb{E}[Z] \ge z$ and $\mathbb{E}[Z] < z$. For the first case, the lower bound automatically follows. Next, we prove the second case. In this case,

$$\mathrm{Var}(Z) = \mathbb{E}[(Z-\mathbb{E}[Z])^2] \ge \mathbb{E}[\mathbb{1}(Z\ge z)(Z-\mathbb{E}[Z])^2] \ge \mathbb{E}[\mathbb{1}(Z\ge z)(z-\mathbb{E}[Z])^2] = \Pr[Z\ge z](z-\mathbb{E}[Z])^2,$$

from which the lower bound follows.

By symmetry, for any $z \in \mathbb{R}$, such that $\Pr[Z < z] \le 1/2$, one can show that

$$z - \sqrt{2\operatorname{Var}(Z)} \le \mathbb{E}[Z] \le z + \sqrt{\frac{\operatorname{Var}(Z)}{\Pr[Z \le z]}}.$$

Since for any z, either $\Pr[Z > z] \le 1/2$ or $\Pr[Z < z] \le 1/2$, Hence, either the first bound in the Lemma or the above bound holds for each z, therefore for any $z \in \mathbb{R}$,

$$z - \max\left\{\sqrt{\frac{\operatorname{Var}(Z)}{\Pr[Z \ge z]}}, \sqrt{2\operatorname{Var}(Z)}\right\} \le \mathbb{E}[Z] \le z + \max\left\{\sqrt{\frac{\operatorname{Var}(Z)}{\Pr[Z \le z]}}, \sqrt{2\operatorname{Var}(Z)}\right\}.$$

The second bound in the lemma is implied by the above bound. □

E.2. Proof of Lemma E.5

Proof of Lemma E.5. The upper bound on $\mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|]$ follows from the Cauchy-Schwarz inequality,

$$\mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|] \le \sqrt{\mathbb{E}_{\mathcal{D}}[|u \cdot x_i^b|^2]} \le \|\Sigma\| \le 1.$$

Next, we prove the lower bound. From Markov's inequality

$$\Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \ge 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]] = \Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^4 \ge 4C^2(\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2])^2] = \frac{\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^4]}{4C^2(\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2])^2} \le \frac{1}{4C},$$

where the last step uses $L4 - L2$ hypercontractivity.

Then, from the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \cdot \mathbb{1}(\|x_i^b \cdot u\|^2 \ge 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2])] &\le \sqrt{\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\left(\|x_i^b \cdot u\|^2 \ge 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]\right)\right] \cdot \mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^4]} \\
&\le \sqrt{\Pr_{\mathcal{D}}\left[\|x_i^b \cdot u\|^2 \ge 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]\right] \cdot C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]^2} \\
&\le \frac{1}{2}\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2].
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \cdot \mathbb{1}(\|x_i^b \cdot u\|^2 < 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2])] &= \mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2] - \mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \cdot \mathbb{1}(\|x_i^b \cdot u\|^2 \ge 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2])] \\
&\ge \frac{1}{2}\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\|x_i^b \cdot u\|^2 \cdot \mathbb{1}\left(\|x_i^b \cdot u\|^2 < 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]\right)\right] \\
&\le \mathbb{E}_{\mathcal{D}}\left[\|x_i^b \cdot u\| \cdot \sqrt{2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]} \cdot \mathbb{1}\left(\|x_i^b \cdot u\|^2 < 2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]\right)\right] \\
&\le \sqrt{2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]} \cdot \mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|].
\end{aligned}$$

Combining the above two equations we get

$$\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|] \ge \frac{\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]}{2\sqrt{2C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]}} = \frac{\sqrt{\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]}}{2\sqrt{2C}}.$$

□

F. Multi-filtering

In this section, we state the subroutine MULTIFILTER, a simple modification of BASICMULTIFILTER algorithm in (Diakonikolas et al., 2020b).

The subroutine takes a weight vector β, a real function z^b on batches, and a parameter θ as input and produces new weight vectors.

This subroutine is used only when:

$$\mathrm{Var}_{B,\beta}(z^b) > c_3 \log^2(2/\alpha)\theta, \tag{28}$$

where c_3 is an universal constant (Same as $2 * C$, where C is the constant in BASICMULTIFILTER algorithm in (Diakonikolas et al., 2020b)).

When the variance of z^b for good batches is smaller than θ and the weight vector β is nice that is $\beta^G \geq 3/4|G|$, then at least one of the new weight vectors produced by this subroutine has a higher fraction of weights in good vector than the original weight vector β.

Algorithm 3 MULTIFILTER

Input: Set B, α, β, $\{z^b\}_{b\in B}$, θ. {Input must satisfy Condition (28)}
Let $a = \inf\{z : \sum_{b:z^b<z} \beta^b \leq \alpha\beta^B/8\}$ and $b = \sup\{z : \sum_{b:z^b>z} \beta^b \leq \alpha\beta^B/8\}$
Let $B' = \{b \in B : z^b \in [a, b]\}$
if $\mathrm{Var}_{B',\beta}(z^b) \leq \frac{c_3 \log^2(2/\alpha)\theta}{2}$ **then**
Let $f^b = \min_{z\in[a,b]} |z^b - z|^2$, and the new weight of each batch $b \in B$ be

$$\beta^b_{\text{new}} = \Big(1 - \frac{f^b}{\max_{b\in B:\beta^b>0} f^b}\Big)\beta^b \tag{29}$$

NEWWEIGHTS $\leftarrow \{\beta_{\text{new}}\}$
else
Find $z \in \mathbb{R}$ and $R > 0$ such that sets $B' = \{b \in B : z^b \geq z - R\}$ and $B'' = \{b \in B : z^b < z + R\}$ satisfy

$$(\beta^{B'})^2 + (\beta^{B''})^2 \leq (\beta^B)^2, \tag{30}$$

and

$$\min\Big(1 - \tfrac{\beta^{B'}}{\beta^B}, 1 - \tfrac{\beta^{B''}}{\beta^B}\Big) \geq \tfrac{48\log(\frac{2}{\alpha})}{R^2}. \tag{31}$$

{Existence of such z and R is guaranteed as shown in Lemma 3.6 of (Diakonikolas et al., 2020b).}
For each $b \in B$, let $\beta^b_1 = \beta^b \cdot \mathbb{1}(b \in B')$ and $\beta^b_2 = \beta^b \cdot \mathbb{1}(b \in B'')$. Let $\beta_1 = \{\beta^b_1\}_{b\in B}$ and $\beta_2 = \{\beta^b_2\}_{b\in B}$.
NEWWEIGHTS $\leftarrow \{\beta_1, \beta_2\}$
end if
Return(NEWWEIGHTS)

In BASICMULTIFILTER subroutine of (Diakonikolas et al., 2020b) input is not restricted by the condition in Equation (28). However, when input meets this condition BASICMULTIFILTER and its modification MULTIFILTER behaves the same.

Therefore, the guarantees for weight vectors returned by MULTIFILTER follows from the guarantees of BASICMULTIFILTER in (Diakonikolas et al., 2020b). We characterize these guarantees in Theorem F.1.

Theorem F.1. *Let $\{z^b\}_{b\in B}$ be collection of real numbers associated with batches, β be a weight vector, and threshold $\theta > 0$ be such that condition in (28) holds. Then* MULTIFILTER*$(B, \beta, \{z^b\}_{b\in B}, \theta_1)$ returns a list* NEWWEIGHTS *containing either one or two new weight vectors such that,*

1. *Sum of square of the total weight of new weight vectors is bounded by the square of the total weight of β, namely*

$$\sum_{\widetilde{\beta}\in \text{NEWWEIGHTS}} (\widetilde{\beta}^B)^2 \leq (\beta^B)^2. \tag{32}$$

2. *In the new weight vectors returned the weight of at least one of the weight vectors has been set to zero, that is for each weight vector* $\widetilde{\beta} \in$ NEWWEIGHTS*,*

$$\{b : \tilde{\beta}^b > 0\} \subset \{b : \beta^b > 0\}, \tag{33}$$

3. *If weight vector* β *is such that* $\beta^G \geq 3|G|/4$ *and for good batches the variance* $Var_G(z^b) \leq \theta$ *is bounded, then for at least one of the weight vector* $\widetilde{\beta} \in$ NEWWEIGHTS*,*

$$\frac{\beta^G - \tilde{\beta}^G}{\beta^G} \leq \frac{\beta^B - \widetilde{\beta}^B}{\beta^B} \cdot \frac{1}{24\log(2/\alpha)}. \tag{34}$$

Proof. When the list NEWWEIGHTS contains one weight vector it is generated using Equation (29), and when the list NEWWEIGHTS contains one weight vector it is generated using Equations (30) and (31). In both cases, item 1 and item 2 of the Theorem follow immediately from these equations. The last item follows from Corollary 3.8 in (Diakonikolas et al., 2020b). □

F.1. Guarantees for the use of MULTIFILTER in Algorithm 1

The following Theorem characterizes the use MULTIFILTER by our algorithm. The proof of the theorem is similar to the proofs for the main algorithm in (Diakonikolas et al., 2020b).

Theorem F.2. *At the end of Algorithm 1 the size of* M *is at most* $4/\alpha^2$ *and the algorithm makes at most* $\mathcal{O}(|B|/\alpha^2)$ *calls to* MULTIFILTER*. And, if for every use of subroutine* MULTIFILTER *by the algorithm we have* $Var_G(z^b) \leq \theta$ *then there is at least one triplet* (β, w, κ) *in* M *such that* $\beta^G \geq 3|G|/4$*.*

Proof. First note that the if blocks in Algorithm 1 ensures that for every use of subroutine MULTIFILTER Equation (28) is satisfied, therefore we can use the guarantees in Theorem F.1.

First we upper bound the size of M.

The progress of Algorithm 1 may be described using a tree. The internal nodes of this tree are the weight vectors that have gone through subroutine MULTIFILTER at some point of the algorithm, and children of these internal nodes are new weight vectors returned by MULTIFILTER. Observe that any weight vector β encountered in Algorithm 1 is ignored iff $\beta^B < \alpha|B|/2$. If it is not ignored then either it is added to M (in form of a triplet), or else it goes through subroutine MULTIFILTER.

It follows that, if a node β is an internal node or a leaf in M then

$$\beta^B \geq \alpha|B|/2. \tag{35}$$

From Equation (32), it follows that the total weight squared for each node is greater than equal to that of its children. It follows that the total weight squared of the root, β_{init} is greater than equal to the sum of the square of weights of all the leaves. And since all weight vectors in M are among the leaves of the tree, and have total weight at least $\alpha|B|/2$,

$$(\beta^B_{\text{init}})^2 \geq \sum_{\beta \in M} (\beta^B)^2 \geq \sum_{\beta \in M} (\frac{\alpha|B|}{2})^2,$$

here the last step follows from Equation (35). Using $\beta^B_{\text{init}} = |B|$, in the above equation we get $|M| \leq 4/\alpha^2$.

Similarly, it can be shown that the number of branches in the tree is at most $\mathcal{O}(1/\alpha^2)$. Item 2 in Theorem F.1 implies that each iteration of MULTIFILTER zeroes out the weight of one of the batches. Hence for any weight β at depth d, we have $\beta^B \leq |B| - d$. Therefore, the depth of the tree can't be more than $|B|$. Hence, the number of nodes in the tree is upper bounded by $\mathcal{O}(|B|/\alpha^2)$. And since each call to MULTIFILTER corresponds to a non-leaf node in the tree, the total calls to MULTIFILTER by Algorithm 1 are upper bounded by $\mathcal{O}(|B|/\alpha^2)$.

Next, we show that if for each use of MULTIFILTER we have $\text{Var}_G(z^b) \leq \theta$ then one of the weight vector $\beta \in M$ must satisfy $\beta^G \geq 3|G|/4$.

Let $\beta_0 = \beta_{\text{init}}$ and suppose for each i, weight vectors β_i and β_{i+1} are related as follows:

$$\frac{\beta_i^G - \beta_{i+1}^G}{\beta_i^G} \leq \frac{\beta_i^B - \beta_{i+1}^B}{\beta_i^B} \cdot \frac{1}{24\log(2/\alpha)}. \tag{36}$$

Then Lemma 3.12 in (Diakonikolas et al., 2020b) showed that under the above relation, for each i, we have $\beta_i^G \geq 3|G|/4$.

We show that there is a branch of the tree such that β_i and β_{i+1} are related using the above equation, where for each i, β_i denote the weight vector corresponding to the node at i^{th} level in this branch. From the preceding discussion, this would imply that for each i, $\beta_i^G \geq 3|G|/4$.

We prove it by induction. For $i = 0$, we select $\beta_i = \beta_{\text{init}}$. Note that $\beta_{\text{init}}^G = |G|$, hence $\beta_i^G \geq 3|G|/4$.

If β_i is a leaf then the branch is complete. Else, since $\beta_i^G \geq 3|G|/4$, item 3 in Theorem F.1 implies that we can select one of the child of β_i as β_{i+1} so that (36) holds. Then from the preceding discussion, we have $\beta_{i+1}^G \geq 3|G|/4$. By repeating this argument, we keep finding the next node in the branch, until we reach the leaf. Next, we argue that the leaf at the end of this branch must be in M.

Let β denote the weight vector for the leaf. From the above discussion, it follows that $\beta^G \geq 3|G|/4$. Hence, $\beta^B \geq \beta^G \geq 3|G|/4 \geq 3\alpha|B|/4 > \alpha|B|/2$.

As discussed earlier any leaf β is not part of M iff $\beta^B \leq \alpha|B|/2$. Hence, the leaf at the end of the above branch must be in M. This concludes the proof of the Theorem. □

G. Eliminating Additional Distributional Assumptions

In this section, we discuss how we can remove assumptions 2 and 5 regarding the distribution of data in Section 2 of the main paper. We demonstrate that our results can still be achieved without these assumptions.

Assumption 2 states that there exists a constant $C_1 > 0$ such that for random samples $(x_i^b, y_i^b) \sim \mathcal{D}$, we have $\|x_i^b\| \leq C_1\sqrt{d}$ almost surely. In the non-batch setting, Cherapanamjeri et al. (2020) (Cherapanamjeri et al., 2020b) have shown that this assumption is not limiting. They have proven that if other assumptions are met, then there exists a constant C_1 such that the probability of $\|x_i^b\| \leq C_1\sqrt{d}$ exceeds 0.99. Thus, discarding the samples where $\|x_i^b\| > C_1\sqrt{d}$ does not significantly reduce the dataset's size. Additionally, it has minimal impact on the covariance matrix and hypercontractivity constants of the distribution This reasoning can be easily extended to the batch setting. In the batch setting, we first exclude samples from batches where $\|x_i^b\| > C_1\sqrt{d}$. We then remove batches that have been reduced by more than 10% of their original size. Since, on average, this operation would remove $\leq 1\%$ of samples from genuine batches, a simple argument using the Markov inequality shows that the probability of removing a genuine batch is at most 10%. It can be demonstrated that with high probability, the fraction of genuine batches that are removed for any component is $\lesssim 10\%$. Therefore, assumption 2 regarding data distribution is not required, and this simple procedure can be used to enforce assumption 2, resulting in a decrease in batch size and α by at most 10%. Consequently, the guarantees in our theorem are altered by only a small factor.

Assumption 5 states that the noise distribution is symmetric. We can address this by employing a simple technique. Let's consider two independent samples (x_i^b, y_i^b) and (x_{i+1}^b, y_{i+1}^b), where $y_j^b = w^* \cdot x_j^b + n_j^b$ for $j \in \{i, i+1\}$. We define $\tilde{x}_i^b = (x_i^b - x_{i+1}^b)/\sqrt{2}$, $\tilde{y}_i^b = (y_i^b - y_{i+1}^b)/\sqrt{2}$, and $\tilde{n}_i^b = (n_i^b - n_{i+1}^b)/\sqrt{2}$. Since n_i^b and n_{i+1}^b are i.i.d., the distribution of $\tilde{n}_i^b$ is symmetric around 0 and the variance of $\tilde{n}_i^b$ matches that of n_i^b. Moreover, the covariance of $\tilde{x}_i^b$ is the same as that of x_i^b, and we have $\tilde{y}_i^b = w^* \cdot \tilde{x}_i^b + \tilde{n}_i^b$. Therefore, the new sample $(\tilde{x}_i^b, \tilde{y}_i^b)$ obtained by combining two i.i.d. samples (x_i^b, y_i^b) and (x_{i+1}^b, y_{i+1}^b) in a batch satisfies the same distributional assumptions as before, and in addition, ensures a symmetric noise distribution. We can apply this approach to combine every two samples in a batch, which only reduces the batch size by a constant factor of 1/2. Thus, the assumption of symmetric noise can be eliminated by increasing the required batch sizes in our theorems by a factor of 2.

H. Proof of Theorem B.1

In Section H.1, we state and prove two auxiliary lemmas that will be used in proving Theorem B.1, and in Section H.2, we prove Theorem B.1.

We will use the following notation in describing the auxiliary lemmas and in the proofs.

Let $S := \{(x_i^b, y_i^b) : b \in G, i \in [n]\}$ denote the collection of all good samples. Note that $|S| = |G|n$.

For any function h over (x, y), we denote the expectation of h w.r.t. uniform distribution on subset $S' \subseteq S$ by $\mathbb{E}_{S'}[h(x_i^b, y_i^b)] := \sum_{(x_i^b, y_i^b) \in S'} \frac{h(x_i^b, y_i^b)}{|S'|}$.

H.1. Auxiliary lemmas

In this subsection, we state and prove Lemmas H.1 and H.2. We will use these lemmas in proof of Theorem B.1 in the following subsection.

In the next lemma, for any unit vectors u, we bound the expected second moment of the tails of $|x_i^b \cdot u|$, for covariate x_i^b of a random sample from the distribution $\mathcal{D}$.

Lemma H.1. *For all $\theta > 1$, and all unit vectors $u \in \mathbb{R}^d$,*

$$\Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta] \leq \frac{1}{\theta^2} \text{ and } \mathbb{E}_{\mathcal{D}}[\mathbb{1}(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta) \cdot \|x_i^b \cdot u\|^2] \leq \frac{\sqrt{C}}{\theta}$$

Proof. The first part of the lemma follows from Markov's inequality,

$$\Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta] = \Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^4 \geq C\theta^2] = \frac{\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^4]}{C\theta^2} \leq \frac{1}{\theta^2},$$

where the last step uses $L4 - L2$ hypercontractivity. This proves the first bound in the lemma.

For the second bound, note that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\mathbb{1}(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta) \cdot \|x_i^b \cdot u\|^2] &\overset{(a)}{\leq} \sqrt{\mathbb{E}_{\mathcal{D}}[\mathbb{1}(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta)] \cdot \mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^4]} \\
&\overset{(b)}{\leq} \sqrt{\Pr_{\mathcal{D}}[\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta] \cdot C\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2]^2} \\
&\overset{(c)}{\leq} \frac{\sqrt{C}}{\theta}\,\mathbb{E}_{\mathcal{D}}[\|x_i^b \cdot u\|^2],
\end{aligned}$$

here (s) follows from the Cauchy-Schwarz inequality, (b) uses $L4 - L2$ hypercontractivity, and (c) follows from the first bound in the lemma. □

In the next lemma, for any unit vectors u, we provide a high probability bound on the expected second moment of the tails of $|x_i^b \cdot u|$, wheres x_i^b are covariates of samples in good batches G.

Lemma H.2. *For any given $\theta > 1$, and $|G|n = \Omega(d\theta^2 \log(\frac{C_1 d\theta}{C}))$, with probability at least $1 - 2/d^2$, for all unit vectors u,*

$$\mathbb{E}_S\Big[\mathbb{1}\Big(\|x_i^b \cdot u\|^2 \geq 3\sqrt{C}\theta\Big) \cdot \|x_i^b \cdot u\|^2\Big] \leq \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right).$$

The following lemma restates Lemma 5.1 of (Cherapanamjeri et al., 2020a). The lemma shows that for any large subset of S, the covariance of covariates x_i^b in S is close to the true covariance for distribution $\mathcal{D}$ of samples. We will use this result in proving Lemma H.2.

Lemma H.3. *For any fix $\theta > 1$, and $|G|n = \Omega(d\theta^2 \log(d\theta))$, with probability at least $1 - 1/d^2$ for all subsets of $S' \subseteq S$ of size $\geq (1 - \frac{1}{\theta^2})|S|$, we have*

$$\Sigma - \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I \preceq \mathbb{E}_{S'}[x_i^b (x_i^b)^\intercal] \preceq \Sigma + \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I.$$

Remark H.1. Lemma 5.1 of (Cherapanamjeri et al., 2020a) assumes that hypercontractive parameter C is a constant and its dependence doesn't appear in their lemma but is implicit in their proof. hides/ignores its dependence.

The following corollary is a simple consequence. We will use this corollary in proving Lemma H.2.

Corollary H.4. *For any fix $\theta > 1$, and $|G|n = \Omega(d\theta^2 \log(d\theta))$, with probability at least $1 - 1/d^2$ for all subsets $S' \subseteq S$ of size $\leq \frac{|S|}{\theta^2}$ and all unit vectors u, we have*

$$\frac{|S'|}{|S|} \cdot \mathbb{E}_{S'}[(x_i^b \cdot u)^2] \preceq \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right).$$

Proof. Consider any set S' of size $\leq \frac{|S|}{\theta^2}$. Since $|S \setminus S'| \geq (1 - \frac{1}{\theta^2})|S|$, applying Lemma H.3 for $S \setminus S'$ and S,

$$\Sigma - \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I \preceq \mathbb{E}_{S \setminus S'}[x_i^b (x_i^b)^\intercal],$$

and

$$\mathbb{E}_S[x_i^b (x_i^b)^\intercal] \preceq \Sigma + \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I.$$

Next,

$$\begin{aligned}
\mathbb{E}_S[x_i^b (x_i^b)^\intercal] &= \frac{|S'|}{|S|}\mathbb{E}_{S'}[x_i^b (x_i^b)^\intercal] + \frac{|S \setminus S'|}{|S|}\mathbb{E}_{S \setminus S'}[x_i^b (x_i^b)^\intercal] \\
\implies |S'|\,\mathbb{E}_{S'}[x_i^b (x_i^b)^\intercal] &= |S|\,\mathbb{E}_S[x_i^b (x_i^b)^\intercal] - |S \setminus S'|\,\mathbb{E}_{S \setminus S'}[x_i^b (x_i^b)^\intercal]).
\end{aligned}$$

Combining the previous three equations,

$$\begin{aligned}
|S'|\,\mathbb{E}_{S'}[x_i^b (x_i^b)^\intercal] &\preceq |S|\left(\Sigma + \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I\right) - |S \setminus S'|\left(\Sigma - \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I\right) \\
&\preceq |S'|\Sigma + (|S| + |S \setminus S'|)\mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I \preceq \frac{1}{\theta^2}|S|\Sigma + 2|S|\mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I \preceq 3|S|\mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right) \cdot I,
\end{aligned}$$

where the last line used $\Sigma \preceq I$, $|S'| \leq |S|/\theta^2$, $C \geq 1$, and $1/\theta^2 \leq 1/\theta$ for $\theta \geq 1$.

Finally, observing that for any unit vector $u^\intercal\,\mathbb{E}_{S'}[x_i^b (x_i^b)^\intercal]u = \mathbb{E}_{S'}[(x_i^b \cdot u)^2]$ completes the proof. □

Now we complete the proof of the Lemma H.2 with help of the above corollary.

Proof of Lemma H.2. From Lemma H.1 we have $\mathbb{E}_{\mathcal{D}}[1(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta)] = \Pr[\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta] \leq \frac{1}{\theta^2}$ Applying Chernoff bound for random variable $\mathbb{1}(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta)$,

$$\Pr\left[\mathbb{E}_S[1(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta)] \leq \frac{2}{\theta^2}\right] = \Pr\left[\frac{1}{|S|}\sum_{(i,b)\in S} 1(\|x_i^b \cdot u\|^2 \geq \sqrt{C}\theta) \leq \frac{2}{\theta^2}\right] \leq \exp\left(-\frac{|S|}{3\theta^2}\right).$$

Hence, for a fix unit vector u, with probability $\geq 1 - \exp\left(-\frac{|S|}{3\theta^2}\right)$

$$\mathbb{E}_S[1(\|x_i^b \cdot u\|^2 \leq \sqrt{C}\theta)] \leq |S|\frac{2}{\theta^2}.$$

Next, we show that this bound holds uniformly over all unit vectors u.

Consider an $\sqrt{\frac{\sqrt{C}\theta}{2C_1 d}}-$ net of unit sphere $\{u \in \mathbb{R}^d : \|u\| \leq 1\}$ such that for any vector u in this ball there exist a u' in the net such that $\|u - u'\| \leq \sqrt{\frac{\sqrt{C}\theta}{2C_1 d}}$. The standard covering argument (Vershynin, 2012) shows the existence of such a net of size $e^{\mathcal{O}(d\log(\frac{C_1 d}{C\theta}))}$. Then from the union bound, for all vectors u in this net with probability at least $1 - e^{\mathcal{O}(d\log(\frac{C_1 d}{C\theta}))}e^{-\frac{|S|}{3\theta^2}}$,

$$\mathbb{E}_S[1(\|x_i^b \cdot u\|^2 \leq \sqrt{C}\theta)] \leq |S|\frac{2}{\theta^2}.$$

Since $\frac{|S|}{3\theta^2} = \frac{|G|n}{3\theta^2} \gg d\log(\frac{C_1 d\theta}{C}) \geq d\log(\frac{C_1 d}{C\theta}))$, therefore, $e^{\mathcal{O}(d\log(\frac{C_1 d}{C\theta}))}e^{-\frac{|S|}{3\theta^2}} \ll e^{-\frac{|S|}{6\theta^2}} \ll 1/d^2$.

Now consider any vector u in unit ball and u' in the net such that $\|u-u'\| \leq \sqrt{\frac{\sqrt{C}\theta}{2C_1 d}}$. Then

$$\begin{aligned}(x_i^b \cdot u)^2 = (x_i^b \cdot (u' + (u-u')))^2 &= 2(x_i^b \cdot u')^2 + (x_i^b \cdot (u-u'))^2 \\ &\leq 2(x_i^b \cdot u')^2 + 2\|u-u'\|^2\|x_i^b\|^2 \\ &\leq 2(x_i^b \cdot u')^2 + 2\frac{\sqrt{C}\theta}{2C_1 d}C_1 d \leq 2(x_i^b \cdot u')^2 + \sqrt{C}\theta,\end{aligned}$$

where in the last line we used the assumption that $\|x_i^b\| \leq C_1\sqrt{d}$. When $(x_i^b \cdot u')^2 \leq \sqrt{C}\theta$, then above sum is bounded by $2\sqrt{C}\theta$. It follows that with probability $\geq 1 - 1/d^2$, for all unit vectors u,

$$\mathbb{E}_S[1(\|x_i^b \cdot u\|^2 \leq 3\sqrt{C}\theta)] \leq |S|\frac{2}{\theta^2}.$$

Applying Corollary H.4 for $S' = \{\|x_i^b \cdot u\|^2 \leq 3\sqrt{C}\theta\}$, proves the lemma

$$\mathbb{E}_S[\mathbb{1}(\|x_i^b \cdot u\|^2 \geq 3\sqrt{C}\theta) \cdot \|x_i^b \cdot u\|^2] = \frac{|S'|}{|S|}\mathbb{E}_{S'}[\|x_i^b \cdot u\|^2] \leq \mathcal{O}\left(\frac{\sqrt{C}}{\theta}\right).$$

□

H.2. Proof of Theorem B.1

Proof of Theorem B.1. Note that

$$\begin{aligned}\mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] &= \frac{1}{|G|}\sum_{b\in G}\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2 \\ &= \frac{1}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}\nabla f_i^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\right)^2 \\ &= \frac{1}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}\big(\nabla f_i^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)\cdot u]\big)\right)^2,\end{aligned}$$

where in the last step we used the expectation of batch and sample gradients are the same, namely $\mathbb{E}_{\mathcal{D}}[\nabla f_i^b(w,\kappa)\cdot u] = \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]$.

For any positive $\rho > 0$ and unit vector u, define

$$g_i^b(w,\kappa,u,\rho) := \frac{\nabla f_i^b(w,\kappa)\cdot u}{\|x_i^b \cdot u\| \vee \rho}\rho.$$

Recall that for a good batch $b \in G$, $y_i^b = w^* \cdot x_i^b + n_i^b$. Using this in equation (3), for any good batch $b \in G$, we have

$$\nabla f_i^b(w,\kappa) = \frac{(x_i^b \cdot (w-w^*) - n_i^b)}{|x_i^b \cdot (w-w^*) - n_i^b| \vee \kappa}\kappa x_i^b. \tag{37}$$

Combining the above two equations,

$$g_i^b(w,\kappa,u,\rho) = \kappa\rho\left(\frac{(x_i^b \cdot (w-w^*) - n_i^b)}{\|x_i^b \cdot (w-w^*) - n_i^b\| \vee \kappa}\right)\left(\frac{x_i^b \cdot u}{\|x_i^b \cdot u\| \vee \rho}\right). \tag{38}$$

From the above expression it follows that $|g_i^b(w,\kappa,u,\rho)| \leq \kappa\rho$ a.s.

We will choose ρ later in the proof. Let

$$Z_i^b(w,\kappa,u,\rho) := g_i^b(w,\kappa,u,\rho) - \mathbb{E}_{\mathcal{D}}\big[g_i^b(w,\kappa,u,\rho)\big].$$

and

$$\begin{aligned}\tilde{Z}_i^b(w,\kappa,u,\rho) &:= \nabla f_i^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}\big[\nabla f_i^b(w,\kappa)\cdot u\big] - Z_i^b(w,\kappa,u,\rho)\\ &= \nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho) - \mathbb{E}_{\mathcal{D}}\big[\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big].\end{aligned}$$

When w, u, κ, and ρ are fixed or clear from the context, we will omit them from the notation of Z_i^b and $\tilde{Z}_i^b$. Then,

$$\begin{aligned}\mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] &= \frac{1}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}(Z_i^b+\tilde{Z}_i^b)\right)^2\\ &\le \frac{2}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}Z_i^b\right)^2 + \frac{2}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}\tilde{Z}_i^b\right)^2\\ &\le \frac{2}{|G|}\sum_{b\in G}\left(\frac{1}{n}\sum_{i\in n}Z_i^b\right)^2 + \frac{2}{|G|}\sum_{b\in G}\frac{1}{n}\sum_{i\in n}(\tilde{Z}_i^b)^2, \qquad (39)\end{aligned}$$

here in the last step we used Jensen's inequality $(\mathbb{E}[Z])^2 \le \mathbb{E}[Z^2]$.

We bound the two summations separately. To bound the first summation we first show that Z_i^b are bounded, and then use Bernstein's inequality. We bound the second term using Lemma H.2 and Lemma H.1.

From (38), it follows that $|g_i^b(w,\kappa,u,\rho)| \le \kappa\rho$ a.s., and therefore, $|Z_i^b| \le 2\kappa\rho$.

Since $|Z_i^b|$ is bounded by $2\kappa\rho$, it is a $(2\kappa\rho)^2$ sub-gaussian random variable. Using the fact that the sum of sub-gaussian random variables is sub-gaussian, the sum $\sum_{i=1}^n Z_i^b$ is $n(2\kappa\rho)^2$ sub-gaussian random variable. Since square of a sub-gaussian is sub-exponential (Philippe, 2015) (Lemma 1.12), hence $(\sum_{i=1}^n Z_i^b)^2 - \mathbb{E}_{\mathcal{D}}(\sum_{i=1}^n Z_i^b)^2$ is sub-exponential with parameter $16n(2\kappa\rho)^2$.

Bernstein's inequality (Philippe, 2015) (Theorem 1.12) for sub-Gaussian random variables implies that with probability $\ge 1-\delta$,

$$\frac{1}{|G|}\sum_{b\in G}\left(\left(\sum_{i=1}^n Z_i^b\right)^2 - \mathbb{E}_{\mathcal{D}}\left[\left(\sum_{i=1}^n Z_i^b\right)^2\right]\right) \le 16n(2\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}.$$

Since Z_i^b are zero mean independent random variables,

$$\mathbb{E}_{\mathcal{D}}\left[\left(\sum_{i=1}^n (Z_i^b)\right)^2\right] = n\,\mathbb{E}_{\mathcal{D}}\big[(Z_i^b)^2\big].$$

We bound the expectation on the right,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\big[(Z_i^b)^2\big] &= \mathbb{E}_{\mathcal{D}}\Big[\big(g_i^b(w,\kappa,u,\rho) - \mathbb{E}_{\mathcal{D}}\big[g_i^b(w,\kappa,u,\rho)\big]\big)^2\Big] \\
&\overset{(a)}{\le} \mathbb{E}_{\mathcal{D}}\Big[\big(g_i^b(w,\kappa,u,\rho)\big)^2\Big] \\
&\overset{(b)}{\le} \mathbb{E}_{\mathcal{D}}[(n_i^b + (w-w^*)\cdot x_i^b)^2 (x_i^b\cdot u)^2] \\
&\overset{(c)}{=} \mathbb{E}_{\mathcal{D}}[(n_i^b)^2 (x_i^b\cdot u)^2] + \mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2 (x_i^b\cdot u)^2] \\
&\overset{(d)}{\le} \mathbb{E}_{\mathcal{D}}[(n_i^b)^2]\,\mathbb{E}_{\mathcal{D}}[(x_i^b\cdot u)^2] + \sqrt{\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^4]\,\mathbb{E}_{\mathcal{D}}[(u\cdot x_i^b)^4]} \\
&\overset{(e)}{\le} \sigma^2\,\mathbb{E}_{\mathcal{D}}[(x_i^b\cdot u)^2] + \sqrt{C^2\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]^2\,\mathbb{E}_{\mathcal{D}}[(u\cdot x_i^b)^2]^2} \\
&\overset{(f)}{\le} \sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2],
\end{aligned}$$

here inequality (a) uses that squared deviation is smaller than mean squared deviation, inequality (b) follows from the definition of g_i^b in (38), inequality (c) follows from the independence of n_i^b and x_i^b, inequality (d) follows the Cauchy–Schwarz inequality, (e) uses the L-4 to L-2 hypercontractivity assumption $\mathbb{E}_{\mathcal{D}}[(u\cdot(x_i^b))^4] \le C$, and (f) follows as for any unit vector $\mathbb{E}_{\mathcal{D}}[(x_i^b\cdot u)^2] \le \|\Sigma\| \le 1$.

Combining the last three equations, we get that with probability $\ge 1-\delta$,

$$\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n Z_i^b\right)^2 \le n(\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]) + 64n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}. \tag{40}$$

The above bound holds for given fixed values of parameters κ, w, and u. To extend the bound for all values of these parameters (for appropriate ranges of interest), we will use the covering argument.

With the help of the covering argument, we show that with probability $\ge 1-\delta e^{\mathcal{O}(d\log(C_1 dn)} - \frac{1}{d^2}$, for all unit vectors u, all vectors w and $\kappa \le (\sigma + \|w-w^*\|)d^2 n$,

$$\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n Z_i^b(w,\kappa,u,\rho)\right)^2 \le \frac{5}{2}\sigma^2 n + 13Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + 384n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}. \tag{41}$$

We delegate the proof of Equation (41) using Equation (40) and the covering argument to the very end. The use of covering argument is rather standard. The main subtlety is that the above bound holds for all vectors w. The cover size of all d dimensional vectors is infinite. To overcome this difficulty we first take union bound for vectors for all w such that $\|w-w^*\| \le R$ for an appropriate choice of R. To extend it to any w for which $\|w-w^*\| > R$ is large we show that the behavior of the above quantity on the left for such a w can be approximated by its behavior for $w' = w^* + (w-w^*)\frac{R}{\|w-w^*\|}$.

Note that dividing Equation (41) by n^2 bounds the first term in Equation (39). Next, we bound the second term in Equation(39). Note that

$$\begin{aligned}
\frac{1}{n|G|}\sum_{b\in G}\sum_{i\in n}(\tilde{Z}_i^b)^2 &\le \frac{1}{n|G|}\sum_{b\in G}\sum_{i\in n}\big(\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho) - \mathbb{E}_{\mathcal{D}}\big[\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big]\big)^2 \\
&\le \frac{2}{n|G|}\sum_{b\in G}\sum_{i\in n}\Big(\big(\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big)^2 + \big(\mathbb{E}_{\mathcal{D}}\big[\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big]\big)^2\Big) \\
&\le \frac{2}{n|G|}\sum_{b\in G}\sum_{i\in n}\Big(\big(\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big)^2 + \mathbb{E}_{\mathcal{D}}\Big[\big(\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big)^2\Big]\Big).
\end{aligned}$$

From the definitions of $g_i^b(w,\kappa,u,\rho)$ and $\nabla f_i^b(w,\kappa)$,

$$\begin{aligned}|\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)| &= \mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|\nabla f_i^b(w,\kappa)\cdot u - \frac{\rho}{\|x_i^b\cdot u\|}\nabla f_i^b(w,\kappa)\cdot u\right| \\ &\leq \mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|\nabla f_i^b(w,\kappa)\cdot u\right| \\ &\leq \kappa\left|x_i^b\cdot u\right|\cdot\mathbb{1}(\|x_i^b\cdot u\|\geq\rho).\end{aligned}$$

From the above equation, it follows that

$$\mathbb{E}_{\mathcal{D}}\Big[\big(\nabla f_i^b(w,\kappa)\cdot u - g_i^b(w,\kappa,u,\rho)\big)^2\Big] \leq \kappa^2\,\mathbb{E}_{\mathcal{D}}\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big].$$

Combining the above three bounds,

$$\begin{aligned}\frac{1}{n|G|}\sum_{b\in G}\sum_{i\in n}(\tilde{Z}_i^b)^2 &\leq \frac{2\kappa^2}{n|G|}\sum_{b\in G}\sum_{i\in n}\Big(\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2 + \mathbb{E}_{\mathcal{D}}\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big]\Big) \\ &= 2\kappa^2\Big(\mathbb{E}_S\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big] + \mathbb{E}_{\mathcal{D}}\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big]\Big),\end{aligned}$$

here the last line uses the fact that S is the collection of all good samples.

For $\rho^2\geq 3\sqrt{C}$, and $|G|n=\Omega(d\rho^4\log(\frac{C_1d\rho}{C}))$, Lemma H.2 implies that with probability at least $1-2/d^2$, for all unit vectors u, we have

$$\mathbb{E}_S\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big] = \mathbb{E}_S\Big[\mathbb{1}(\|x_i^b\cdot u\|^2\geq\rho^2)\left|x_i^b\cdot u\right|^2\Big] \leq \mathcal{O}(\sqrt{C}/\rho^2).$$

And from Lemma H.1, for $\rho^2\geq\sqrt{C}$ and any unit vectors u,

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{1}(\|x_i^b\cdot u\|\geq\rho)\left|x_i^b\cdot u\right|^2\Big] \leq \mathcal{O}(\sqrt{C}/\rho^2).$$

By combining the above three bounds it follows that, if $\rho^2\geq\sqrt{C}$, and $|G|n=\Omega(d\rho^4\log(\frac{C_1d\rho}{C}))$, with probability at least $1-2/d^2$, for all unit vectors u,

$$\frac{1}{n|G|}\sum_{b\in G}\sum_{i\in n}(\tilde{Z}_i^b(w,\kappa,u,\beta))^2 \leq \mathcal{O}\left(\frac{\sqrt{C}\kappa^2}{\rho^2}\right).$$

Combining the above bound, Equation (41) and (39) we get that if $\rho^2=\Omega(\sqrt{C})$, and $|G|=\Omega(\frac{d\rho^4}{n}\log(\frac{C_1d\rho}{C}))$ then with probability $\geq 1-\delta e^{\mathcal{O}(d\log(C_1dn)}-\frac{3}{d^2}$, for all unit vectors u, all vectors w and $\kappa\leq(\sigma+\|w-w^*\|)d^2n$,

$$\begin{aligned}&\mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] \\ &\leq \frac{2}{n^2}\left(\frac{5}{2}\sigma^2n + 13Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + 384n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|},\sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}\right) + \mathcal{O}\left(\frac{\sqrt{C}\kappa^2}{\rho^2}\right).\end{aligned}$$

Recall that $1\leq\mu_{\max}\leq\frac{d^4n^2}{C}$. Choose $\rho^2=\mu_{\max}\sqrt{C}n$. Note that $\sqrt{\mu_{\max}(\sigma^2+C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2])}\leq(\sigma+\|w-w^*\|)d^2n$. Then from the above equation choosing $\rho^2=\mu_{\max}\sqrt{C}n$, for all

$$\kappa\leq\sqrt{\mu_{\max}(\sigma^2+C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2])},$$

with probability $\geq 1-\delta e^{\mathcal{O}(d\log(C_1dn)}-\frac{3}{d^2}$, for all unit vectors u, all vectors w ,

$$\begin{aligned}&\mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] \\ &\leq \mathcal{O}\left(\frac{\sigma^2+C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n}\right)\left(1+n\mu_{\max}^2\sqrt{C}\max\left\{\frac{2\ln(1/\delta)}{|G|},\sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}\right).\end{aligned}$$

Choose $\delta = e^{-\Theta(d\log(C_1 dn)}$, and $|G| = \Omega(\frac{d\rho^4}{n}\log(\frac{C_1 d\rho}{C}) + C\mu_{\max}^4 dn^2\log(C_1 dn)) = \Omega(\rho_{\max}^4 n^2 d\log(d))$. Then with probability $\geq 1 - \frac{4}{d^2}$, for all unit vectors u, all vectors w and for all $\kappa^2 \leq \mu_{\max}(\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2])$,

$$\mathbb{E}_G\Big[\big(\nabla f^b(w,\kappa)\cdot u - \mathbb{E}_{\mathcal{D}}[\nabla f^b(w,\kappa)\cdot u]\big)^2\Big] \leq \mathcal{O}\bigg(\frac{\sigma^2 + C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2]}{n}\bigg),$$

which is the desired bound.

We complete the proof by proving Equation (41).

Proof of Equation (41) To complete the proof of the theorem next we prove Equation (41) with the help of Equation (40) and covering argument. To use the covering argument, we first show that $g_i^b(w,\kappa,u,\rho)$ do not change by much by slight deviation of these parameters. From the definition of $Z_i^b(w,\kappa,u,\rho)$, the same conclusion would then hold for it.

By the triangle inequality,

$$\begin{aligned}
&|g_i^b(w,\kappa,u,\rho) - g_i^b(w',\kappa',u',\rho)| \\
&\leq |g_i^b(w',\kappa',u,\rho) - g_i^b(w',\kappa',u',\rho)| + |g_i^b(w,\kappa',u,\rho) - g_i^b(w',\kappa',u,\rho)| + |g_i^b(w,\kappa,u,\rho) - g_i^b(w,\kappa',u,\rho)|.
\end{aligned}$$

We bound each term on the right one by one. To bound these terms we use Equation (38), the assumption that $\|x_i^b\| \leq C_1\sqrt{d}$ and the definition of the function $g()$. For the first term,

$$|g_i^b(w',\kappa',u,\rho) - g_i^b(w',\kappa',u',\rho)| \leq \|(u-u')x_i^b\|\kappa' \leq C_1\|u-u'\|\sqrt{d}\kappa',$$

for the second term,

$$|g_i^b(w,\kappa',u,\rho) - g_i^b(w',\kappa',u,\rho)| \leq |u\cdot x_i^b|\cdot|(w-w')\cdot x_i^b| \leq \|x_i^b\|^2\cdot\|w-w'\| \leq C_1^2 d\|w-w'\|,$$

and for the last term

$$|g_i^b(w,\kappa,u,\rho) - g_i^b(w,\kappa',u,\rho)| \leq |\kappa-\kappa'|\cdot|u\cdot x_i^b| \leq C_1\sqrt{d}|\kappa-\kappa'|.$$

Combining the three bounds,

$$|g_i^b(w,\kappa,u,\rho) - g_i^b(w',\kappa',u',\rho)| \leq C_1\|u-u'\|\sqrt{d}\kappa' + C_1^2 d\|w-w'\| + C_1\sqrt{d}|\kappa-\kappa'|.$$

For $\|u-u'\| \leq 1/(24C_1 d^5 n^3)$, $\kappa' \leq 2d^4\sigma n^2$, $\|w-w'\| \leq \sigma/(12dC_1^2 n)$ and $|\kappa-\kappa'| \leq \sigma/(12C_1 dn)$,

$$|g_i^b(w,\kappa,u,\rho) - g_i^b(w',\kappa',u',\rho)| \leq \sigma/4n \text{ a.s.}$$

This would imply,

$$|Z_i^b(w,\kappa,u,\rho) - Z_i^b(w',\kappa',u',\rho)| \leq \sigma/2n \text{ a.s.}$$

Using this bound,

$$\begin{aligned}
\frac{1}{|G|}\sum_{b\in G}\bigg(\sum_{i=1}^n Z_i^b(w,\kappa,u,\rho)\bigg)^2 &\leq \frac{1}{|G|}\sum_{b\in G}\bigg(\sum_{i=1}^n\Big(Z_i^b(w',\kappa',u',\rho') + \frac{\sigma}{2n}\Big)\bigg)^2 \\
&\leq \frac{2}{|G|}\sum_{b\in G}\bigg(\sum_{i=1}^n Z_i^b(w',\kappa',u',\rho')\bigg)^2 + \frac{2}{|G|}\sum_{b\in G}\bigg(\sum_{i=1}^n\frac{\sigma}{2n}\bigg)^2 \\
&\leq \frac{2}{|G|}\sum_{b\in G}\bigg(\sum_{i=1}^n Z_i^b(w',\kappa',u',\rho')\bigg)^2 + \frac{\sigma^2}{2}. \qquad (42)
\end{aligned}$$

Let $\mathcal{U} := \{u\in\mathbb{R}^d : \|u\| = 1\}$, $\mathcal{W} := \{w\in\mathbb{R}^d : \|w-w^*\| \leq d^2\sigma n)\}$, and $\mathcal{K} := \big[0, 2d^4\sigma n^2\big]$.

Standard covering argument shows that there exist covers such that

$$\mathcal{U}' \subseteq \mathcal{U} : \forall u \in \mathcal{U}, \min_{u' \in \mathcal{U}'} \|u - u'\| \leq \frac{1}{(24C_1 d^5 n^3)}, \tag{43}$$

$$\mathcal{W}' \subseteq \mathcal{W} : \forall w \in \mathcal{W}, \min_{w' \in \mathcal{W}'} \|w - w'\| \leq \frac{\sigma}{12C_1^2 dn}, \tag{44}$$

and

$$\mathcal{K}' \subseteq \mathcal{K} : \forall \kappa \in \mathcal{K}, \min_{\kappa' \in \mathcal{K}', \kappa' \geq \kappa} |\kappa - \kappa'| \leq \frac{\sigma}{12C_1 dn}, \tag{45}$$

and the size of each is $|\mathcal{U}'|, |\mathcal{W}'|, |\mathcal{K}'| \leq e^{\mathcal{O}(d \log(C_1 dn)}$.

In equation (40), taking the union bound over all elements in $\mathcal{U}'$, $\mathcal{W}'$ and $\mathcal{K}'$, it follows that with probability $\geq 1 - \delta e^{\mathcal{O}(d \log(C_1 dn)}$, for all $u' \in \mathcal{U}'$, $w' \in \mathcal{W}'$ and $\kappa' \in \mathcal{K}'$

$$\frac{1}{|G|} \sum_{b \in G} \left(\sum_{i=1}^{n} Z_i^b(w', \kappa', u', \rho) \right)^2 \leq n(\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[((w' - w^*) \cdot x_i^b)^2]) + 64n(\kappa'\rho)^2 \max\left\{ \frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}} \right\}.$$

Combining the above bound with Equation (42), it follows that with probability $\geq 1 - \delta e^{\mathcal{O}(d \log(C_1 dn)}$, for all $u \in \mathcal{U}$, $w \in \mathcal{W}$ and $\kappa \in \mathcal{K}$ and elements u', w' and κ' in the respective nets satisfying equations (43),(44), and (45),

$$\begin{aligned}
&\frac{1}{|G|} \sum_{b \in G} \left(\sum_{i=1}^{n} Z_i^b(w, \kappa, u, \rho) \right)^2 \leq 2n(\sigma^2 + C\, \mathbb{E}_{\mathcal{D}}[((w' - w^*) \cdot x_i^b)^2]) + 128n(\kappa'\rho)^2 \max\left\{ \frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}} \right\} + \frac{\sigma^2}{2} \\
&\leq 2n(\sigma^2(1 + \frac{1}{4n}) + C\, \mathbb{E}_{\mathcal{D}}[((w' - w^*) \cdot x_i^b)^2]) + 128n(\kappa\rho)^2 \max\left\{ \frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}} \right\} \\
&\leq 2n(\frac{5}{4}\sigma^2 + 2C\, \mathbb{E}_{\mathcal{D}}[((w - w^*) \cdot x_i^b)^2]) + 128n(\kappa\rho)^2 \max\left\{ \frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}} \right\}.
\end{aligned} \tag{46}$$

here (a) follows from the bound $\kappa \geq \kappa'$ in Equation (45), and (b) follows by first writing $w' - w^* = (w - w^*) + (w' - w)$ and then using the bound $\|w' - w\| \leq \frac{\sigma}{12C_1^2 dn}$ in Equation (44).

Next, we further remove the restriction $w \in \mathcal{W}$ and extend the above bound to all vectors w.

Consider a $w \notin \mathcal{W}$ and $\kappa \in \left[0, (\sigma + \|w - w^*\|)d^2 n\right]$. From the definition of $\mathcal{W}$, we have $\|w - w^*\| > d^2\sigma n$. Let $w' = w^* + \frac{w - w^*}{\|w - w^*\|} d^2\sigma n$ and $\kappa' = \frac{d^2\sigma n}{\|w - w^*\|}\kappa$. Observe that $\|w' - w\| = d^2\sigma n$ and

$$\kappa' \leq (\sigma + \|w - w^*\|)d^2 n \frac{d^2\sigma n}{\|w - w^*\|} \leq d^2\sigma n \frac{d^2\sigma n}{\|w - w^*\|} + d^4\sigma n^2 \leq d^2\sigma n + d^4\sigma n^2 \leq 2d^4\sigma n^2,$$

hence, $w' \in \mathcal{W}$ and $\kappa' \in \mathcal{K}$. From Equation (38),

$$\begin{aligned}
&\left|\frac{\|w-w^*\|}{d^2\sigma n} g_i^b(w',\kappa',u,\rho) - g_i^b(w,\kappa,u,\rho)\right| \\
&\overset{(a)}{=} \frac{\rho\|x_i^b\cdot u\|}{\|x_i^b\cdot u\|\vee\rho}\cdot\left|\frac{\|w-w^*\|}{d^2\sigma n}\kappa'\left(\frac{(x_i^b\cdot(w'-w^*)-n_i^b)}{\|x_i^b\cdot(w'-w^*)-n_i^b\|\vee\kappa'}\right) - \kappa\left(\frac{(x_i^b\cdot(w-w^*)-n_i^b)}{\|x_i^b\cdot(w-w^*)-n_i^b\|\vee\kappa}\right)\right| \\
&\overset{(b)}{\le} \|x_i^b\cdot u\|\cdot\left|\kappa\left(\frac{(x_i^b\cdot\frac{w-w^*}{\|w-w^*\|}d^2\sigma n - n_i^b)}{\|x_i^b\cdot\frac{w-w^*}{\|w-w^*\|}d^2\sigma n - n_i^b\|\vee\frac{d^2\sigma n}{\|w-w^*\|}\kappa}\right) - \kappa\left(\frac{(x_i^b\cdot(w-w^*)-n_i^b)}{\|x_i^b\cdot(w-w^*)-n_i^b\|\vee\kappa}\right)\right| \\
&= \|x_i^b\cdot u\|\cdot\left|\kappa\left(\frac{(x_i^b\cdot(w-w^*)-\frac{d^2\sigma n}{\|w-w^*\|}n_i^b)}{\|x_i^b\cdot(w-w^*)-\frac{d^2\sigma n}{\|w-w^*\|}n_i^b\|\vee\kappa}\right) - \kappa\left(\frac{(x_i^b\cdot(w-w^*)-n_i^b)}{\|x_i^b\cdot(w-w^*)-n_i^b\|\vee\kappa}\right)\right| \\
&\overset{(c)}{\le} \|x_i^b\cdot u\|\cdot\left|\frac{d^2\sigma n}{\|w-w^*\|}n_i^b - n_i^b\right| \\
&\overset{(d)}{\le} C_1\sqrt{d}|n_i^b|\frac{d^2\sigma n}{\|w-w^*\|} \\
&\overset{(e)}{\le} C_1\sqrt{d}|n_i^b|,
\end{aligned}$$

here (a) follows from the definition of g_i^b, inequality (b) follows as $\rho \le \|x_i^b\cdot u\|\vee\rho$, inequality (c) uses the fact that for any a, Δ and $b \ge 0$, we have $|b\frac{a+\Delta}{(a+\Delta)\vee b} - b\frac{a}{a\vee b}| \le |\Delta|$, inequality (c) uses $\|x_i^b\| \le C_1\sqrt{d}$ and the last inequality (e) uses $\|w-w^*\| > d^2\sigma n$.

Therefore,

$$\left|\frac{\|w-w^*\|}{d^2\sigma n} Z_i^b(w',\kappa',u,\rho) - Z_i^b(w,\kappa,u,\rho)\right| \le C_1\sqrt{d}\big(|n_i^b| + \mathbb{E}[|n_i^b|]\big) \le C_1\sqrt{d}\big(|n_i^b| + \sigma\big).$$

From the above equation,

$$\begin{aligned}
&\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n Z_i^b(w,\kappa,u,\rho)\right)^2 \le \frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n\left(\frac{\|w-w^*\|}{d^2\sigma n}Z_i^b(w',\kappa',u,\rho) + C_1\sqrt{d}|n_i^b| + C_1\sqrt{d}\sigma\right)\right)^2 \\
&\overset{(a)}{\le} \frac{1}{|G|}\sum_{b\in G}\left(3\left(\sum_{i=1}^n\frac{\|w-w^*\|}{d^2\sigma n}Z_i^b(w',\kappa',u,\rho)\right)^2 + 3\left(\sum_{i=1}^n C_1\sqrt{d}|n_i^b|\right)^2 + 3\left(\sum_{i=1}^n C_1\sqrt{d}\sigma\right)^2\right) \\
&\overset{(b)}{\le} \frac{3\|w-w^*\|^2}{d^4\sigma^2n^2}\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n Z_i^b(w',\kappa',u,\rho)\right)^2 + \frac{3dC_1^2}{|G|}\sum_{b\in G}\left(n\sum_{i=1}^n|n_i^b|^2\right) + 3dC_1^2n^2\sigma^2 \\
&\overset{(c)}{\le} \frac{3\|w-w^*\|^2}{d^4\sigma^2n^2}\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^n Z_i^b(w',\kappa',u,\rho)\right)^2 + 3dC_1^2n^2\sigma^2(d^2+1),
\end{aligned}$$

here (a) and (b) uses $(\sum_{i=1}^t z_i) \le t\sum_{i=1}^t z_i^2$ and inequality (c) holds with with probability $\ge 1 - \frac{1}{d^2}$ by Markov inequality, as $Pr[\frac{1}{n|G|}\sum_{b\in G}\sum_{i=1}^n|n_i^b|^2 > d^2\,\mathbb{E}_{\mathcal{D}}[(n_i^b)^2)] \le \frac{1}{d^2}$.

Recall that $\|w'-w\| \le d^2\sigma n$ and $\kappa \le 2d^4\sigma n^2$, therefore in the above equation, we can bound the first term on the right by

using high probability bound in Equation (46). Then, with probability $\geq 1 - \delta e^{\mathcal{O}(d\log(C_1 dn)} - \frac{1}{d^2}$,

$$\begin{aligned}
&\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^{n} Z_i^b(w,\kappa,u,\rho)\right)^2 \\
&\leq \frac{3\|w-w^*\|^2}{d^4\sigma^2 n^2}\left(2n(\frac{5}{4}\sigma^2 + 2C\,\mathbb{E}_{\mathcal{D}}[((w'-w^*)\cdot x_i^b)^2]) + 128n(\kappa'\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}\right) \\
&\quad + 3dC_1^2 n^2\sigma^2(d^2+1) \\
&\overset{(a)}{=} 12Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + \frac{15\|w-w^*\|^2}{2d^4\sigma^2 n^2}n\sigma^2 + 384n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\} + 3dC_1^2 n^2\sigma^2(d^2+1) \\
&\overset{(b)}{\leq} 13Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + 384n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\},
\end{aligned}$$

here equality (a) uses the relation $w' - w^* = \frac{w-w^*}{\|w-w^*\|}d^2\sigma n$ and $\kappa' = \frac{d^2\sigma n}{\|w-w^*\|}\kappa$, and (b) follows as $C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] \geq C\frac{\|w-w^*\|^2\|\Sigma\|}{C_3} = C\frac{\|w-w^*\|^2}{C_3} \geq Cd^4\sigma^2 n^2/C_3$, where C_3 is condition number of Σ, hence $C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] \gg \frac{15\|w-w^*\|^2}{2d^4\sigma^2 n^2}n\sigma^2$ and $C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] \geq C\frac{\|w-w^*\|^2}{C_3} \gg \frac{15\|w-w^*\|^2}{2d^4\sigma^2 n^2}n\sigma^2$ and $C\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] \gg 3dC_1^2 n^2\sigma^2(d^2+1)$.

The above bound holds for all unit vectors u, $w \notin \mathcal{W}$ and $\kappa \leq (\sigma + \|w-w^*\|)d^2 n$,

$$\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^{n} Z_i^b(w,\kappa,u,\rho)\right)^2 \leq 13Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + 450n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}.$$

Recall that bound in Equation (46) holds for all unit vectors u, $w \in \mathcal{W}$ and $\kappa \leq \mathcal{K}'$ with probability $\geq 1 - \delta e^{\mathcal{O}(d\log(C_1 dn)}$. Note that for $w \in \mathcal{W}$, $(\sigma + \|w-w^*\|)d^2 n \leq 2d^4\sigma n^2$, hence $[0, (\sigma + \|w-w^*\|)d^2 n] \subseteq \mathcal{K}'$. Hence the above bound holds for all unit vectors u, $w \in \mathcal{W}$ and $\kappa \leq (\sigma + \|w-w^*\|)d^2 n$. Combining the two bounds, with probability $\geq 1 - \delta e^{\mathcal{O}(d\log(C_1 dn)} - \frac{1}{d^2}$, for all unit vectors u, all vectors w and $\kappa \leq (\sigma + \|w-w^*\|)d^2 n$,

$$\frac{1}{|G|}\sum_{b\in G}\left(\sum_{i=1}^{n} Z_i^b(w,\kappa,u,\rho)\right)^2 \leq \frac{5}{2}\sigma^2 n + 13Cn\,\mathbb{E}_{\mathcal{D}}[((w-w^*)\cdot x_i^b)^2] + 384n(\kappa\rho)^2\max\left\{\frac{2\ln(1/\delta)}{|G|}, \sqrt{\frac{2\ln(1/\delta)}{|G|}}\right\}.$$

This completes the proof of Equation (41). □