Optimal Robust Linear Regression in Nearly Linear Time

Yeshwanth Cherapanamjeri∗ Efe Aras† Nilesh Tripuraneni‡ Michael I. Jordan§
Nicolas Flammarion¶ Peter L. Bartlett‖

July 17, 2020

Abstract

We study the problem of high-dimensional robust linear regression where a learner is given access to \(n\) samples from the generative model \(Y = \langle X, w^* \rangle + \epsilon\) (with \(X \in \mathbb{R}^d\) and \(\epsilon\) independent), in which an \(\eta\) fraction of the samples have been adversarially corrupted. We propose estimators for this problem under two settings: (i) \(X\) is L4-L2 hypercontractive, \(\mathbb{E}[XX^\top]\) has bounded condition number and \(\epsilon\) has bounded variance and (ii) \(X\) is sub-Gaussian with identity second moment and \(\epsilon\) is sub-Gaussian. In both settings, our estimators:

1. Achieve optimal sample complexities and recovery guarantees up to log factors and
2. Run in near linear time (\(\tilde{O}(nd/\eta^6)\)).

Prior to our work, polynomial time algorithms achieving near optimal sample complexities were only known in the setting where \(X\) is Gaussian with identity covariance and \(\epsilon\) is Gaussian, and no linear time estimators were known for robust linear regression in any setting. Our estimators and their analysis leverage recent developments in the construction of faster algorithms for robust mean estimation to improve runtimes, and refined concentration of measure arguments alongside Gaussian rounding techniques to improve statistical sample complexities.

1 Introduction

Least-squares regression is amongst the oldest and most fundamental statistical methods – its use dates back over two centuries to the seminal works of Gauss [1809] and Legendre [1805] who used it to estimate the trajectories of celestial bodies. Since then, least-squares regression has found numerous applications in varied fields like finance [Dielman, 2001], epidemiology [Bhaskaran et al., 2013], astronomy [Isobe et al., 1990] and biostatistics [McDonald, 2009]. However, the algorithmic primitives underlying these estimators assume that the data being used to perform such analysis is clean and well-curated. This assumption is simply not true in modern datasets which often contain extremely noisy and sometimes even adversarial data.

In line with such considerations, we study the problem of high-dimensional robust linear regression. In the standard linear regression setup, one has access to \(n\) i.i.d. samples from the following generative model
\begin{equation*}
Y = \langle X, w^* \rangle + \epsilon \text{ where } X \text{ and } \epsilon \text{ are independent and } \epsilon \text{ is mean 0 with } \text{Var}(\epsilon) \leq \sigma^2. \text{ In the robust setting, an adversary is allowed to observe the generated samples and arbitrarily perturb an } \eta \text{ fraction of them. The goal is to estimate the true regression vector, } w^*, \text{ given such corrupted samples. We consider two concrete scenarios: }
\end{equation*}

1. **Heavy Tailed Case:** \( X \) satisfies an L4-L2 hyper-contractivity assumption and

2. **Sub-Gaussian Case:** \( X \) and \( \epsilon \) are sub-Gaussian with \( \mathbb{E}[XX^\top] = I \).

In both settings, our proposed estimators run in near linear-time with near optimal sample complexity and achieve information theoretically optimal recovery guarantees of \( O(\sigma \sqrt{n}) \) in the heavy tailed case assuming a bound on the condition number of \( \mathbb{E}[XX^\top] \) and \( O(\sigma \eta \log 1/\eta) \) in the sub-Gaussian case which is information theoretically optimal up to a factor \( O(\sqrt{\log 1/\eta}) \). In contrast, all previous approaches to robust linear regression suffer from either sub-optimal sample complexities, slow running times or both [Klivans et al., 2018, Prasad et al., 2018, Diakonikolas et al., 2019a, b].

Our algorithm is based on the iterative robust gradient estimation framework from [Diakonikolas et al., 2019a] and [Prasad et al., 2018]. In both these works, a candidate parameter vector is maintained and a robust estimate of the gradient of the function being optimized is obtained by running a robust estimation procedure on a collection of contaminated gradient estimates. For example, the gradients at some current estimate \( w \) would consist of the vectors \( G_i(w) = (\langle X_i, w \rangle - Y_i)X_i \) for linear regression. The robust gradient is then used to iteratively improve the estimate \( w \). Since, the number of iterations is typically small, the computational cost of the algorithm is dominated by the cost of a single iteration. To obtain our improved computational performance, we leverage recent developments in algorithms for robust mean estimation [Cheng et al., 2019, Dong et al., 2019] to implement the robust estimation step in near-linear time.

To ensure the success of the algorithm, one needs to show that the robust estimation procedure returns a good estimate of the gradient. In typical approaches to this problem, one is required to exhibit a set of weights, \( s_i \), on the gradients such that \( \sum s_i G_i(w) \) is close to the true gradient and \( \sum s_i G_i(w)G_i(w)^\top \) is spectrally bounded. The usual approach to ensuring this condition is to bound the spectral norm of the random tensor \( W = n^{-1} \sum X_i^\otimes 4 \). However, note that the length of a typical \( X_i \) is on the order of \( \sqrt{d} \) which means that the spectral norm of \( X_i^\otimes 4 \) is \( O(d^2) \). Therefore, one would need \( n \approx d^2 \) for the spectral norm of \( W \) to be small. To circumvent this issue, we instead work with a carefully chosen subset of our data based on our current estimate and use an intricate generalization of Gaussian rounding schemes previously employed in the context of heavy-tailed mean estimation [Lecué and Depersin, 2019] to exhibit the existence of a suitable set of weights without sacrificing sample complexity.

For the sub-Gaussian case, we generalize results from [Diakonikolas et al., 2019b] for the case where the covariates \( X \) and error \( \epsilon \) are Gaussian to the milder assumptions of sub-Gaussianity while still achieving the same recovery error. Interestingly, we show that one can obtain this improved error using the heavy tailed algorithm by adding a single correction step at its conclusion to improve the estimate from an error of \( O(\sqrt{n}\sigma) \) to \( O(\eta \log 1/\eta\sigma) \). This is also applicable to the robust mean estimation scenario where it conceptually simplifies the presentation of [Cheng et al., 2019] and [Dong et al., 2019].

### 1.1 Related work

There is long line of work in the statistics community which has led to the construction of estimators robust to the presence of adversarial noise. For example, for the problem of robust mean estimation, Huber [1964] first proposed an estimator in the one dimensional case which was later generalized to the multi-dimensional case by Tukey [1975]. Since then, the expansion of these ideas has resulted in the
development of robust estimators for other tasks like linear regression and covariance estimation [Huber, 2004]. Unfortunately, there are no known polynomial time algorithms to compute several of these estimators. As a remedy to these shortcomings, a recent line of work in the computer science community have devised polynomial time robust estimators achieving near optimal recovery guarantees for a range of statistical estimation problems including mean estimation, linear regression and covariance estimation [Lai et al., 2016, Diakonikolas et al., 2016, 2017, Prasad et al., 2018, Diakonikolas et al., 2019a, Klivans et al., 2018, Kothari et al., 2018, Diakonikolas et al., 2018a, Hopkins and Li, 2018, Steinhardt et al., 2018, Charikar et al., 2017, Diakonikolas et al., 2018b, Diakonikolas and Kane, 2019, Diakonikolas et al., 2019b]. For robust mean estimation, recent work has resulted in algorithms running in near linear time and achieving optimal sample complexity and recovery error [Cheng et al., 2019, Dong et al., 2019].

The problem of robust linear regression has been previously considered in the works of Klivans et al. [2018], Diakonikolas et al. [2019b], Prasad et al. [2018], Diakonikolas et al. [2019a], Balakrishnan et al. [2017], Bhatia et al. [2015, 2017], Suggala et al. [2019]. Of these, [Bhatia et al., 2015, 2017, Suggala et al., 2019] only tolerate adversarial corruption in the responses $Y_i$ and do not apply in our scenario where the covariates may also be adversarially manipulated. From the viewpoint of computational complexity, none of the aforementioned works have algorithms with run times sub-quadratic in dimension. In terms of sample complexity, all previous algorithms are sub-optimal requiring the number of samples to be at least quadratic in the dimension barring the work of Diakonikolas et al. [2019b] which applies only in the setting where $X$ and $\epsilon$ are normally distributed.

Our work is most closely related to [Diakonikolas et al., 2019a, Prasad et al., 2018] which consider a general set-up where one aims to optimize a loss function, say $f(w) = \mathbb{E}[(\langle X_i, w \rangle - Y_i)^2]$ for linear regression given access to stochastic gradients from $f$, a fraction of which have been adversarially corrupted. However, while their algorithms may be augmented with faster robust gradient estimation solvers to speed up runtime, their sample complexities are sub-optimal with sample complexities quintic and quadratic in dimension respectively [Diakonikolas et al., 2019a, Prasad et al., 2018] and achieve optimal recovery error only in the heavy tailed setting. Diakonikolas et al. [2019b] achieve near optimal recovery error and sample complexity under the assumption that both $X$ and $\epsilon$ are Gaussian with $X$ having identity covariance matrix but tolerate arbitrary covariance matrices by exploiting a robust covariance estimation procedure for Gaussian inputs at the cost of sub-optimal sample complexity.

Recent Work: Two contemporary works [Bakshi and Prasad, 2020, Zhu et al., 2020] also study the problem of robust linear regression in greater generality than in this paper and in particular, Bakshi and Prasad [2020] obtain nuanced information theoretically optimal recovery guarantees which scale with the number of available moments of the distribution. However, algorithmically exploiting these assumptions is computationally and statistically expensive. Consequently, both these approaches suffer from impractically large runtime ($O(n^{O(k)})$) and sample complexity ($O(d^{O(k)})$)\(^1\). In addition, we also do not require the more restrictive assumption of certifiability on the moments of the distribution.

1.2 Notation

We often use $\mathcal{D}$ to refer to a distribution over $(X,Y)$, $\mathcal{G} = \{G_i\}_{i=1}^n$ and $\mathcal{Z} = \{Z_i\}_{i=1}^N$ to refer to sets of points and $\mathcal{H}$ and its variants to refer to events in a probability space. We use $\Delta_\delta$ to refer to the subset of the probability simplex defined by $\{s \in \mathbb{R}^n : \sum_{i=1}^n s_i = 1, 0 \leq s_i \leq \frac{1}{(1-\delta) n}\}$ and $\mathcal{E}_\delta$ to refer to the extreme points

\(^1\)Assuming $k$ moments of the distribution are available. In our settings, $k \geq 4$. 


of $\Delta_\delta$; that is $E_\delta = \{ s \in \mathbb{R}^n : \sum_{i=1}^n s_i = 1, s_i \in \{ \frac{1}{1-\delta}, 0 \} \}$. Given a set of points $\{Z_i\}_{i=1}^n$, a subset $S \subseteq [n]$ and $s \in \Delta_\delta$, we denote by $E_S[f(X_i)] = |S|^{-1} \sum_{i \in S} f(X_i)$ and $E_s[f(X_i)] = \sum_{i=1}^n s_i f(X_i)$. We use $\text{Supp}(s)$ to denote the set of indices for which $s_i > 0$. We also use $G_i(w) = \langle X_i, w - w^* \rangle - Y_i X_i$ to refer to the gradient of the linear regression objective evaluated on the $i$th sample at $w$, and $G^*(w) = E[G_i(w)]$ to refer to the population gradient at $w$. Finally, for a vector $v$ and matrix $M$, we will use $\|v\|$ and $\|M\|$ to denote the Euclidean and spectral norms of $v$ and $M$ respectively and $\lambda_{\text{max}}(M)$ for the largest eigenvalue of $M$.

## 2 Main Results

In this section, we formally present the main results of our paper. For a generative model, $\mathcal{D}$, we formally describe the corruption model below:

**Definition 2.1.** Given $\eta > 0$ and a distribution, $\mathcal{D}$, a set of samples, $\{(X_i, Y_i)\}_{i=1}^n$, is $\eta$-corrupted if it is generated according to the following model:

1. $\{(X'_i, Y'_i)\}_{i=1}^n$ are generated i.i.d. from the distribution $\mathcal{D}$
2. An adversary is allowed to inspect the samples and arbitrarily perturb any $\eta n$ of them.

For the heavy tailed scenario, we make the following assumptions on the generative model $Y = \langle X, w^* \rangle + \epsilon$:

**Assumption 2.2.** The (uncorrupted) generative model $Y = \langle X, w^* \rangle + \epsilon$ satisfies:

1. $X$ is a random vector with second moment matrix $\Sigma$. Furthermore, $X$ is L4–L2 hyper-contractive. That is, for all $\|u\| = 1$, we have:

$$E[\langle u, X \rangle^4] \leq \tau \cdot \left( E[\langle u, X \rangle^2] \right)^2$$

for an absolute constant $\tau$. For normalization purposes, we assume $\|\Sigma\| = 1$.

2. The condition number of $\Sigma$ is bounded by $\kappa < \infty$.

3. We have that $X$ satisfies $\|X\| \leq C_1 \sqrt{d}$ almost surely for some absolute constant $C_1 > 1$.

4. The noise variable $\epsilon$ is zero mean and independent of $X$ with variance bounded by $\sigma^2$.

Additionally, we assume that $\eta$ satisfies $\eta \leq \frac{C_1}{\kappa^2}$ for a suitably small constant\(^\text{2}\) $C_2 > 0$.

The assumptions stated above are standard for heavy tailed linear regression where some condition on the anti-concentration of $XX^\top$ is required to obtain finite sample complexities [Lecué and Mendelson, 2016, Oliveira, 2016]. Note that the boundedness assumption $\|X\| \leq O(\sqrt{d})$ is not a restriction and is made for ease of presentation – since an unbounded random vector $X$ satisfying hypercontractivity can be truncated at level $O(\sqrt{d})$ without significantly distorting the second moment matrix, $\Sigma$ or L4-L2 hypercontraction constant, $\tau$ (See Lemma B.1 in Appendix B). Given a set of random vectors satisfying Assumption 2.2 save for the boundedness condition in 3., we can simply discard samples with large norms as a preprocessing step so they satisfy Assumption 2.2 without affecting our analysis or conclusions.

The main result of our paper for the heavy tailed scenario is described in the following theorem:

\(^\text{2}\)This condition is explained in Lemma 4.2.
Theorem 2.3. Suppose a distribution $D$ satisfies Assumption 2.2. Then there exists an algorithm which when given $n = \tilde{O}(d/\eta)$ $\eta$-corrupted samples from $D$, runs in time $\tilde{O}(kn d/\eta^6)$ and returns an estimate, $\hat{w}$ satisfying:

$$\|\hat{w} - w^*\| \leq O(\kappa \sqrt{\eta} \sigma),$$

with probability at least 0.9.

In the setting where $\kappa = O(1)$, we obtain the information theoretically optimal recovery error of $O(\sqrt{\eta} \sigma)$ with optimal sample complexity up to log factors (Theorem D.1). Note that it is information theoretically impossible to achieve parameter recovery independent of $\kappa$ under these assumptions (Corollary D.2). For the sub-Gaussian scenario, we make the following stronger assumptions on the generative model:

Assumption 2.4. The generative model $Y = \langle X, w^* \rangle + \epsilon$ satisfies:

1. $X$ is a random vector with second moment matrix $I$. Furthermore, we assume that there exists an absolute constant, $\psi$, such that for every $\|u\| = 1$:

$$P(|\langle X, u \rangle - \mathbb{E}\langle X, u \rangle| \geq t) \leq 2 \exp \left( -\frac{t^2}{2\psi^2} \right).$$

2. The noise variable $\epsilon$ is zero mean, variance $\sigma^2$ independent of $X$ and is sub-gaussian with sub-gaussianity parameter $\phi = O(\sigma)$. That is, $\epsilon$ satisfies:

$$P(|\epsilon| > t) \leq 2 \exp \left( -\frac{t^2}{2\phi^2} \right).$$

Additionally, we assume that $\eta$ satisfies $\eta \leq C_2$ for a sufficiently small constant $C_2 > 0$.

With these stronger assumptions on the generative model, we obtain improved recovery guarantees as detailed in the following theorem:

Theorem 2.5. Suppose a distribution $D$ satisfies Assumption 2.4. Then, there exists an algorithm which when given $n = \tilde{O}(d/\eta^2)$ $\eta$-corrupted samples from $D$, runs in time $\tilde{O}(nd/\eta^6)$ and returns an estimate, $\hat{w}$, satisfying:

$$\|\hat{w} - w^*\| \leq O(\eta \log 1/\eta \sigma),$$

with probability at least 2/3.

We note that in comparison to the heavy tailed scenario, we obtain an improved recovery guarantee of $O(\eta \log 1/\eta \sigma)$ as opposed to the $O(\sqrt{\eta} \sigma)$ error obtained in Theorem 2.3. The guarantee obtained in Theorem 2.5 is information theoretically optimal up to a factor of $\sqrt{\log 1/\eta}$ (Theorem D.3) and the sample complexity is optimal up to log factors [Diakonikolas et al., 2017, Gao, 2017].

Remark: It is possible to improve the runtime dependence on $\eta$ to a logarithmic dependence on $1/\eta$ by appealing to refined solvers [Dong et al., 2019] for the class of semidefinite programs we use here. However, we do not pursue this avenue in this work.
3 Algorithm

In this section, we describe our algorithms which are used in Theorems 2.3 and 2.5. In Subsection 3.1, we describe the algorithm for the heavy tailed scenario and build on this algorithm to construct an algorithm for the sub-Gaussian case in Subsection 3.2. For both algorithms, an important computational primitive we make use of at several points is the following semidefinite program given a set of points \( Z = \{ Z_i \}_{i=1}^n \) and \( \delta > 0 \):

\[
\min_{s \in \Delta_\delta} \lambda_{\text{max}} \left( \sum_{i=1}^n s_i Z_i Z_i^T \right). \tag{MT}
\]

In what follows, we use \( \text{MT}(Z, \delta) \) to denote the program \( \text{MT} \) instantiated with inputs \( Z \) and \( \delta \). Semidefinite programs of the form featured in the above display were used to build near linear time algorithms for robust mean estimation [Cheng et al., 2019] by reducing the problem to a packing/covering SDP and subsequently exploiting fast positive semidefinite programming solvers of Peng et al. [2012]. The dual solution to the above program is then used to refine an estimate of the mean. In contrast, our approach only utilizes primal solutions to the above SDP allowing us to simplify the reduction to packing/covering SDP (See Appendix E) resulting in faster solvers for a set of vectors \( Z \) of any scale as opposed to [Cheng et al., 2019] where \( \|Z_i\| \leq O(\sqrt{d}) \).

3.1 Algorithm for Heavy Tailed Robust Linear Regression

In the heavy tailed setting, our approach builds on the iterative robust gradient estimation framework from [Prasad et al., 2018, Diakonikolas et al., 2019a] where in each step, we construct a robust estimate of the gradient of the function \( f(w) = \mathbb{E}[(Y - \langle w, X \rangle)^2] \). In each iteration, \( t \), we compute the gradient at each of the sample points and then use a solution to \( \text{MT} \) to obtain a reliable estimate of the gradient of \( f \). We then use this gradient to improve our estimate until convergence.

Note that in the non-adversarial scenario, the average of the sample gradients, \( G_t(w_t) = (\langle X_i, w_t \rangle - Y_i)X_i \) provides a good estimate of the population gradient, defined to be \( G^*(w) = \Sigma (w - w^*) \). However, in the robust setting this is no longer true. As a remedy to such issues, \( \text{MT} \), employs a set of weights which allows one to exclude samples which adversely affect the value of the estimated gradient. The main observation is that to effect a mild change in the mean of the sample gradients, an outlier must have an outsized effect on the sample second moment leading to lower weight being assigned to the outlier. Our algorithm is formally described in Algorithm 1 and our gradient estimation procedure is described in Algorithm 2 where we simply use the primal weights from a solution to \( \text{MT} \) to obtain our gradient estimate.

**Algorithm 1 Heavy Tailed Linear Regression**

1: **Input:** Set of sample points \( Z = \{(X_i, Y_i)\}_{i=1}^n \), Outlier Fraction \( \eta \), Length of Parameter Vector \( \|w^*\| \)

2: \( T \leftarrow O(k \log \frac{\|w^*\|}{\eta \sigma \kappa}) \)

3: \( w_0 \leftarrow 0 \)

4: **for** \( t = 0 : T \) **do**

5: \( g_t \leftarrow \text{Gradient Estimation}(Z, \eta, w_t) \)

6: \( w_{t+1} \leftarrow w_t - g_t \)

7: **end for**

8: **Return:** \( w_{T+1} \)
Algorithm 2 Gradient Estimation

1: **Input:** Set of sample points $Z = \{(X_i, Y_i)\}_{i=1}^n$, Outlier Fraction $\eta$, Current Estimate $w$
2: $G = \{G_i(w) = (\langle X_i, w \rangle - Y_i)X_i\}$
3: $s \leftarrow (1 + 5\eta)$-approximate solution to $MT(G, 10\eta)$
4: $g \leftarrow \mathbb{E}_s[G_i(w)]$
5: **Return:** $g$

3.2 Algorithm for Sub-Gaussian Robust Linear Regression

In this subsection, we present our algorithm for robust linear regression in the setting where the covariates are assumed to be sub-Gaussian with identity covariance (see Assumption 2.4). The estimate $w^\dagger$ approximates $w^*$ up to an error of $O(\sqrt{\eta}\sigma)$. Interestingly, we further establish that refining the estimate to $O(\eta \log 1/\eta)$ simply requires solving $MT$ one more time and using the average of sample gradients as a correction factor thus obtaining our refined rates. The procedure is detailed in Algorithm 3. Notably, this procedure can also be used to simplify algorithms for robust mean estimation under sub-Gaussian assumptions.

Algorithm 3 Sub-gaussian Robust Linear Regression

1: **Input:** Set of sample points $Z = \{(X_i, Y_i)\}_{i=1}^n$, Outlier Fraction $\eta$, Length of Parameter Vector $||w^*||$
2: $n_1 \leftarrow \tilde{O}(d/\eta)$
3: $Z_1 \leftarrow$ Random $n_1$ samples from $Z$
4: $\tilde{Z}_1 \leftarrow \{(X_i, Y_i) \in Z_1 : ||X_i|| \leq \tilde{O}(\sqrt{d})\}$
5: $w^\dagger \leftarrow$ Heavy Tailed Linear Regression($\tilde{Z}_1, \eta, ||w^*||$)
6: $Z_2 \leftarrow \{(X_i, Y_i - \langle X_i, w^\dagger \rangle) : (X_i, Y_i) \in Z \setminus Z_1\}$
7: $G = \{G_i = Y_iX_i : (X_i, Y_i) \in Z_2\}$
8: $s \leftarrow (1 + 3\eta)$-approximate solution to $MT(G, 6\eta)$
9: $\hat{w} \leftarrow w^\dagger + \mathbb{E}[G_i]$
10: **Return:** $\hat{w}$

4 Analysis under Deterministic Assumptions

In this section we provide the analysis of our method under deterministic assumptions on the distribution of the good samples.

4.1 Analysis for Heavy Tailed Robust Linear Regression

We prove here the success of our approach in the heavy tailed case. The main steps involved in the proof are the following:

1. **Gradient Estimation:** We show that the vector $G$ output by Algorithm 2 is a good estimate of the true gradient $G^*(w)$. See Lemma 4.2.
2. **Gradient Descent:** We prove that the gradient descent algorithm eventually converges to a good approximation of the regression vector $w^*$. See Lemma 4.3.

To prove the correctness of these two steps, we require the following condition to hold:
Assumption 4.1. For every \( w \in \mathbb{R}^d \), there exists \( S \subset [n] \) such that \( |S| \geq (1 - 5\eta)n \) satisfying:

\[
\|\mathbb{E}_S [G_i(w)] - G^*(w)\| \leq O(\sqrt{\eta}(\|w - w^*\| + \sigma)) \quad \text{and} \quad \|\mathbb{E}_S [G_i(w)G_i^T]\| \leq O(\|w - w^*\|^2 + \sigma^2)
\]

Deriving the concentration tools to prove these conditions is the object of Section 5 where it is shown they uniformly hold, over all \( w \in \mathbb{R}^d \), with probability at least 0.95 (see Lemma 5.5).

4.1.1 Gradient estimation

First we prove the correctness of the Gradient Estimation step of Algorithm 2.

Lemma 4.2. Under Assumption 4.1, there is a universal constant \( C_3 > 0 \) such that for any \( w \in \mathbb{R}^d \), any \((1 + 5\eta)\)-approximate solution, \( s \), to \( \text{MT}(G, 10\eta) \) for \( G = \{G_i(w) = (X_i, w) - Y_iX_i\}_{i=1}^n \), satisfies:

\[
\|\mathbb{E}_s [G_i(w)] - G^*(w)\| \leq C_3 \sqrt{\eta}(\|w - w^*\| + \sigma).
\]

Proof: This result follows from the simple idea that two distributions with small TV distance and bounded second moments have close means. Consider any \( \tilde{s} \in \Delta_{5\eta} \). Applying the Cauchy-Schwarz inequality for any direction \( \|v\| = 1 \) and a coupling argument between \( s \) and \( \tilde{s} \), we get:

\[
\langle \mathbb{E}_s [G_i] - \mathbb{E}_{\tilde{s}} [G_i], v \rangle \leq \sqrt{\mathbb{E}[(G_s - G_{\tilde{s}}, v)^2]} \sqrt{\text{Dist}_{TV}(s, \tilde{s})},
\]

where \( G_s \) denotes the random vector \( G_s = G_i \) with probability \( s_i \) and \( G_s, G_{\tilde{s}} \) are coupled. We first note that

\[
\mathbb{E}[\langle G_s - G_{\tilde{s}}, v \rangle^2] \leq 2\mathbb{E}[\langle G_s, v \rangle^2] + 2\mathbb{E}[\langle G_{\tilde{s}}, v \rangle^2] \leq (1 + 5\eta)\text{OPT}_{5\eta}^* + 2\mathbb{E}[\langle G_{\tilde{s}}, v \rangle^2] \leq 14\mathbb{E}[\langle G_{\tilde{s}}, v \rangle^2].
\]

since \( s \) approximately solves the SDP and \( \tilde{s} \) is feasible for this SDP. A direct application of Lemma F.8 yields \( \text{Dist}_{TV}(s, \tilde{s}) \leq 15\eta \). Therefore, multiplying the two bounds and maximizing over all directions \( \|v\| = 1 \) yields

\[
\|\mathbb{E}_s [G_i] - \mathbb{E}_{\tilde{s}} [G_i]\| = O\left(\sqrt{\eta \lambda_{\max}\left(\mathbb{E}_s[G_iG_i^T]\right)}\right).
\]

Finally using the triangle inequality, we obtain

\[
\|\mathbb{E}_s [G_i] - G^*\| = O\left(\sqrt{\eta \lambda_{\max}\left(\mathbb{E}_s[G_iG_i^T]\right)} + \|\mathbb{E}_s[G_i] - G^*\|\right).
\]

We conclude by applying this result to the weight vector \( \tilde{s} \) defined as \( \tilde{s} = \frac{1}{|S|} \), where \( S \) is the set given in Assumption 4.1. \( \square \)

4.1.2 Gradient Descent

Here we show that the gradient descent (i.e. Algorithm 1) can produce an accurate estimate of the underlying parameter vector, \( w^* \), when used with the gradient estimates of Algorithm 2. The proof amounts to analyzing gradient descent dynamics with deterministic noise in the gradients.

Lemma 4.3. Suppose that Assumptions 4.1 and 2.2 hold. Assuming that \( 1/\kappa > 2C_3\sqrt{\eta} \), Algorithm 1 outputs an estimate \( w_T \) satisfying,

\[
\|w_T - w^*\| \leq O(\kappa\sigma\sqrt{\eta}),
\]

in \( T = O\left(\kappa \log \left(\frac{\|w^*-w_0\|}{C_3\sqrt{\eta}\sigma \kappa}\right)\right) \) iterations, where \( C_3 \) is the same constant as in Lemma 4.2.
Proof. Noting $G^*(w) = \Sigma(w - w^*)$, we consider the evolution of the sequence $w_{t+1} - w^*$:

$$w_t - w^* = w_{t-1} - w^* - \mathbb{E}_t[G_t(w_{t-1})] = (I - \Sigma)(w_{t-1} - w^*) - e_{t-1},$$

where the error sequence $e_t = \mathbb{E}_t[G_t(w_t)] - G^*(w_t)$ uniformly satisfies $\|e_t\| \leq \sqrt{\eta}C_3 \cdot (\|w_t - w^*\| + \sigma)$ from Lemma 4.2. Taking norms of the equation and applying the triangle inequality shows that,

$$\|w_t - w^*\| \leq (\|I - \Sigma\| + \sqrt{\eta}C_3)\|w_{t-1} - w^*\| + \sqrt{\eta}C_3\sigma.$$

Unrolling the recursive inequality shows that,

$$\|w_t - w^*\| \leq (\|I - \Sigma\| + \sqrt{\eta}C_3)^t(\|w_0 - w^*\|) + \sqrt{\eta}C_3\sigma \sum_{i=0}^{t-1} (\|I - \Sigma\| + \sqrt{\eta}C_3)^i.$$

Recall that we normalize $\|\Sigma\| = 1$ and hence, $\|I - \Sigma\| + \sqrt{\eta}C_3 \leq 1 - (1 - \sqrt{\eta}C_3) \leq 1 - \frac{1}{\sqrt{\eta}} \leq e^{-1/2k} < 1$ since we assume that $1/\kappa > 2C_3\sqrt{\eta}$. Using this choice of step-size and summing the geometric series (to $\infty$) implies that,

$$\|w_t - w^*\| \leq e^{-t/(2k)}\|w_0 - w^*\| + 4C_3\sqrt{\eta}\sigma^2.$$

Hence choosing $t = O(\kappa \log \frac{\|w_0 - w^*\|}{C_3\sqrt{\eta}\sigma^2})$ ensures that the upper bound is $O(\kappa\sigma\sqrt{\eta})$. \qed

4.2 Analysis for Sub-Gaussian Robust Linear Regression

From Lemma 4.3, we can assume that we have an initial estimate $w^\dagger$ such that $\|w^\dagger - w^*\| \leq O(\sqrt{\eta})$. Therefore, by subtracting out $(X_i, w^\dagger)$ from all the data points, our problem reduces to a setting where we additionally have $\|w^\dagger\| \leq O(\sqrt{\eta})$. We make this formal in the following assumption:

Assumption 4.4. We assume that there exists a constant $\nu$, such that $\|w^\dagger\| \leq \nu \sigma \sqrt{\eta}$.

To avoid dealing with the randomness of clean samples, we require in addition the following condition:

Assumption 4.5. There exists a set $S \subset [n]$ with $|S| \geq (1 - 3\eta)n$ such that for any $T \subset S$ with $|T| \geq (1 - 10\eta)n$:

$$\|\mathbb{E}_T[Y_iX_i] - w^*\| \leq O(1) \cdot \sigma\eta \log 1/\eta \text{ and } \|\mathbb{E}_T[Y_i^2X_iX_i^T] - \sigma^2 \cdot I\| \leq O(1) \cdot \sigma^2 \eta \log^2 1/\eta.$$

These conditions are shown to hold in Appendix C. We present now the following lemma which states that solving (MT) one more time is enough to refine the estimation rate to $O(\eta \log(1/\eta))$.

Lemma 4.6. Suppose Assumptions 4.4 and 4.5 hold for $Z = \{(X_i, Y_i)\}_{i=1}^n$. Let $G = \{G_i = Y_iX_i : (X_i, Y_i) \in Z\}$. Then, a $(1 + 3\eta)$-approximate solution to $MT(G, 6\eta)$, $s$, satisfies:

$$\|\mathbb{E}_s[G_i] - w^*\| \leq O(1) \cdot \sigma\eta \log 1/\eta.$$

Proof. First, using Assumption 4.5 we easily see that

$$\lambda_{\max} (\mathbb{E}_s[G_iG_i^T]) \leq \sigma^2 + O(1) \cdot \sigma^2 \eta \log^2 1/\eta.$$

Then we decompose $s$ as $s = \sum_{\alpha \in \mathcal{E}_0} \gamma_\alpha \alpha$ by Lemma F.7 with $\sum_{\alpha \in \mathcal{E}_0} \gamma_\alpha = 1$. In addition, we define $\alpha_T = \sum_{i \in T} \alpha_i$. Using the triangle inequality we have:

$$\|\mathbb{E}_s[G_i] - w^*\| \leq \| \sum_{i \in S} s_i(G_i - w^*) \| + \| \sum_{i \in S} s_i(G_i - w^*) \|.$$
We bound the first term as follows:
\[
\left\| \sum_{i \in S} s_{i}(G_{i} - w^*) \right\| \leq \sum_{\alpha \in \mathcal{E}_{\eta}} \gamma_{\alpha} \left\| \sum_{i \in S} \alpha_{i}(G_{i} - w^*) \right\| = \sum_{\alpha \in \mathcal{E}_{\eta}} \gamma_{\alpha} \cdot \frac{1}{\alpha S} \left\| \mathbb{E}_{\text{Supp}(\alpha) \cap S}(G_{i} - w^*) \right\| \leq O(1) \cdot \sigma \eta \log 1/\eta
\]
where the last inequality follows because we have \( \text{Supp}(\alpha) \cap [n] \geq (1 - 10\eta)n \) and from Assumption 4.5. For the second term, we have for all \( \|u\| = 1 \):
\[
\mathbb{E}_{\alpha}[u, (G_{i} - w^*)] \mathbf{1}\{i \notin S\} \leq (\mathbb{E}_{\alpha}[u, G_{i} - w^*])^{2} \mathbf{1}\{i \notin S\})^{1/2}(\mathbb{E}_{\alpha}[\mathbf{1}\{i \notin S\})^{1/2} \leq 2\sqrt{\mathbb{E}_{\alpha}[u, G_{i} - w^*]^{2}} \mathbf{1}\{i \notin S\})^{1/2} \leq 4\sqrt{\mathbb{E}_{\alpha}[u, G_{i}^{2}] \mathbf{1}\{i \notin S\})^{1/2} + 4\sqrt{\mathbb{E}_{\alpha}[u, w^*]^{2}} \mathbf{1}\{i \notin S\})^{1/2}.
\]
We finally bound the last terms in the above display as follows:
\[
\mathbb{E}_{\alpha}[u, G_{i}^{2}] \mathbf{1}\{i \notin S\}) = \mathbb{E}_{\alpha}[u, G_{i}^{2}] \mathbf{1}\{i \notin S\}) \leq (1 + O(1)\eta \log^{2} 1/\eta)\sigma^{2} - \sum_{\alpha \in \mathcal{E}_{\eta}} \gamma_{\alpha} \cdot \alpha S \cdot \mathbb{E}_{\text{Supp}(\alpha) \cap S}[u, G_{i} - w^*]^{2} \mathbf{1}\{i \notin S\}) \leq \sigma^{2} \left(1 + O(1)\eta \log^{2} 1/\eta - (1 - O(1) \cdot \eta \log^{2} 1/\eta)\right) \leq O(1) \cdot \sigma^{2} \eta \log^{2} 1/\eta,
\]
where the last inequality follows from Lemma F.7 and Assumption 4.5. And
\[
\mathbb{E}_{\alpha}[u, w^*]^{2} \mathbf{1}\{i \notin S\}) \leq \|w_{s}\|^{2}\mathbb{E}_{\alpha}[\mathbf{1}\{i \notin S\}) \leq 4\eta \|w_{s}\|^{2} \leq 4\eta^{3/2},
\]
where the last inequality follows from Assumption 4.4. Substituting the above bounds in the previous equations, we conclude the proof of the lemma. \(\square\)

5 Concentration

In this section, we sketch the main arguments establishing the deterministic conditions in Assumption 4.1 required for the success of Algorithm 1. Since, the corresponding proofs for the sub-Gaussian setting are conceptually and technically simpler, we defer them to Appendix C. Informally, under our assumptions, we are required to show for any estimate, \( w \), the existence of a good set \( S \subseteq [n] \) exhibiting strong concentration properties for \( \mathbb{E}_{S}[G_{i}(w)] \) and \( \mathbb{E}_{S}[G_{i}(w)G_{i}(w)^{T}] \). Before we begin, we first note that the following decomposition is useful for the subsequent analysis:
\[
\mathbb{E}_{S}[G_{i}(w)] = (\mathbb{E}_{S}[X_{i}X_{i}^{T}]) (w - w^*) - \mathbb{E}_{S}[\epsilon_{i}X_{i}] \quad \text{and} \quad (1)
\]
\[
\mathbb{E}_{S}[G_{i}(w)G_{i}(w)^{T}] = \mathbb{E}_{S}[(X_{i}, (w - w^*))^{2}]X_{i}X_{i}^{T} - 2\epsilon_{i}(X_{i}, (w - w^*))X_{i}X_{i}^{T} + \epsilon_{i}^{2}X_{i}X_{i}^{T} \quad \text{(2)}
\]

We now establish control on the first term in Eq. (1) through the following lemma which shows that the empirical second moment \( \mathbb{E}_{S}[X_{i}X_{i}^{T}] \) is close to the true second moment matrix for all suitably large sets, \( S \).³

³The set \( S \) is eventually chosen to exclude points corrupted by the adversary.
Lemma 5.1. Let \( \{X_i\}_{i=1}^n \) satisfy Assumption 2.2. Then, there exists a universal constant \( c \) such that if \( n \geq c \log(\frac{d}{\eta^2}) \), with probability at least \( 1 - \frac{1}{d^2} \) for any set of \( (1 - 10\eta)n \) samples, \( S \), we have:

\[
\Sigma - O(\sqrt{n}) \cdot I \leq \mathbb{E}_S[X_iX_i^\top] \leq \Sigma + O(\sqrt{n}) \cdot I.
\]

- Showing the upper bound is straightforward – we simply note that \( \sum_{i \in S} X_iX_i^\top \leq \sum_{i=1}^n X_iX_i \) and that \( \frac{|S|}{n} = 1 + O(\eta) \) – hence we can simply apply the matrix Bernstein inequality to the latter sum to show the result uniformly over all sets \( S \).

- The proof of the lower bound proceeds by first noting that it suffices to show that for any \( S \), \( \mathbb{E}_S[Z_i] \geq u^\top \Sigma u - O(\sqrt{n}) \) where \( Z_i = \langle X_i, u \rangle^2 \) for any \( u \) in an \( \eta \)-net of resolution \( O(\frac{\sqrt{n}}{d}) \), since at this scale the error of replacing \( u \in S^{d-1} \) by its closest approximant in the net is negligible. Second, using the Bernstein inequality, for a fixed direction \( u \), we argue that with high probability at least \( 10\eta \) points lie in the interval \( (q, \infty) \), where \( q \) is defined as \( \mathbb{E}(Z_i \in (q, \infty)) = 20\eta \). Then we establish a bound on the following random variable \( Z = \mathbb{E}[\mathbb{E}[Z_i \mathbb{1}\{Z_i \leq q\}]] \). From our hypercontractivity assumptions, we get \( \mathbb{E}[Z_i \mathbb{1}\{Z_i \leq q\}] \geq u^\top \Sigma u - O(\sqrt{n}) \). We exploit our hypercontractivity assumptions again to establish that \( q \leq O(1/\sqrt{n}) \). As a consequence, we may apply the Bernstein inequality to the random variable \( Z \) to obtain \( Z \geq u^\top \Sigma u - O(\sqrt{n}) \) w.h.p. and since the elements, \( Z_i \in (q, \infty) \) contribute maximally to \( \mathbb{E}[\mathbb{E}[Z_i]] \), the previous expression establishes a lower bound on \( \mathbb{E}_S[Z_i] \) as at least \( 10\eta \alpha \) of the \( Z_i \) are in \( (q, \infty) \). Union-bounding over these high-probability events and the \( \eta \)-net concludes the argument.

To finish controlling \( 1 \) only bounding the term \( \mathbb{E}_S[\epsilon_i X_i] \) remains. In the heavy-tailed setting, it suffices to use the boundedness of the covariance and exploit the randomness in \( \epsilon_i \) to exhibit the result. This allows us to obtain a bound of \( O(\sqrt{n} \sigma) \) for \( \mathbb{E}_S[\epsilon_i X_i] \) for all suitable large \( S \). Since these results follow from straightforward generalizations of concentration techniques for robust mean estimation, we defer these statements and their proofs to Appendix B (See Lemma B.2).

We now state the main lemma to establish Eq. (2), which controls the leading term. Control of the \( \epsilon_i^2 X_iX_i^\top \) term follow from similar techniques as Lemma 5.1 and the \( 2\epsilon_i \langle X_i, u \rangle X_iX_i^\top \) term follows from the Cauchy-Schwarz inequality.

Lemma 5.2. Let \( \mathcal{D} \) satisfy Assumption 2.2. Then, for \( n = \tilde{O}(d/\eta) \) samples from \( \mathcal{D} \), we have with probability at least \( 1 - \frac{1}{d} \) that for every \( ||u|| = 1 \), there exists \( S \subset [n] \) with \( |S| \geq (1 - \eta/6)n \) such that:

\[
\mathbb{E}_S[\langle X_i, u \rangle^2 X_iX_i^\top] \leq C_S I,
\]

and furthermore, for all \( i \in S \), we have \( |\langle X_i, u \rangle| \leq 40/\eta \).

Establishing this result requires several steps,

- First by exploiting convex duality we can see that:

\[
\min_{\alpha \in \Delta_U} \max_{M \geq 0, ||M||=1} \left( \sum_{i \in S} \alpha_i \langle X_i, u \rangle^2 X_iX_i^\top, M \right) = \max_{M \geq 0, ||M||=1} \min_{\alpha \in \Delta_U} \left( \sum_{i \in S} \alpha_i \langle X_i, u \rangle^2 X_iX_i^\top, M \right).
\]  

Apriori the weights \( \alpha_i \) may depend on the vector \( u \) and can be chosen to discard terms from the sum for which \( \langle X_i, u \rangle^2 \) is large. Exchanging the min and max allows \( \alpha \) to also depend on the matrix \( M \) and admits the possibility they may be chosen to discard terms for which the matrix \( M \) has strong overlap.
with $X_i$. The price of exchanging the min and max, is that in the left-side of Eq. (3) the optimal $M$ is always achieved at a rank-one matrix; however in order to appeal to convex duality we must lift the program to an SDP which in general results in a semi-definite matrix $M$ appearing on the right-hand side.

- As a stepping stone we consider the simpler case where $M = vv^T$ is rank-1 and exhibit a set of weights $\alpha_i$ (which may depend $M$) which can avoid $X_i$ which are strongly correlated with $u$ and $v$. This is formally stated in:

**Lemma 5.3.** Let $\mathcal{D}$ satisfy Assumption 2.2. Then, for $n = \tilde{O}(d/\eta)$ samples from $\mathcal{D}$, we have with probability at least $1 - 1/d$, that for all $\|u\| \leq 1$ and $\|v\| \leq 1$:

$$\mathbb{E}_{\eta}[\sum_{i=1}^{n} (X_i, u)^2 (X_i, v)^2 1\{\|X_i, u\| \leq 800/\sqrt{\eta} \land \|X_i, v\| \leq 800/\sqrt{\eta}\}] \leq C_4 \quad \text{and}$$

$$\mathbb{E}_{\alpha}[\mathbb{1}\{\|X_i, u\| \leq 40/\sqrt{\eta} \land \|X_i, v\| \leq 40/\sqrt{\eta}\}] \geq (1 - \eta/100),$$

for some absolute constant $C_4$.

We first argue that given pair $u$ and $v$, we may discard problematic terms in the sum for which $X_i$ overlaps strongly with $u$ or $v$ without changing the number of non-zero terms in the sum by more than a factor of $O(\eta n)$. This allows us to focus our attention on terms satisfying $\|X_i, u\|, \|X_i, v\| \leq O(1/\sqrt{\eta})$.

To obtain the above condition uniformly over all $u$ and $v$, we use a sufficiently fine grid over $(u, v)$ such that approximation error is negligible and apply a union bound over all the elements in the grid.

- The final and most technical step is to demonstrate that the previous argument which is restricted to rank-one matrices $M = vv^T$ can be lifted to arbitrary normalized p.s.d. matrices. We formally state this in:

**Lemma 5.4.** Let $\mathcal{D}$ satisfy Assumption 2.2. Then, for $n = \tilde{O}(d/\eta)$ samples from $\mathcal{D}$, we have with probability at least $1 - 1/d$, that for every $\|u\| = 1$ and p.s.d. matrix $M$ with $\text{Tr} M = 1$:

$$\mathbb{E}_{\eta}[\langle X_i, u \rangle^2 \langle X_i, X_i^T, M \rangle \mathbb{1}\{\|X_i, u\| \leq 40/\sqrt{\eta} \land \langle X_i, X_i^T, M \rangle \leq 400^2/\eta\}] \leq C_6 \quad \text{and}$$

$$\mathbb{E}_{\alpha}[\mathbb{1}\{\|X_i, u\| \leq 40/\sqrt{\eta} \land \langle X_i, X_i^T, M \rangle \leq 400^2/\eta\}] \geq (1 - \eta/25),$$

for some absolute constant $C_6$.

The argument to do this uses an intricate Gaussian rounding argument which generalizes arguments used previously in the context of heavy tailed mean estimation [Lecué and Depersin, 2019] where a rounding argument was used to obtain large deviation bounds of the points $X_i$ in the “direction” $M$. However, these results do not suffice for our application as these only provide us a bound of $O(1/\eta^2)$ as opposed to a constant. We generalize these arguments to obtain variance bounds on $X_i$ in the “direction” $M$. The full details of the proof are presented in Appendix B.

Combining these results establishes Lemma 5.2. Finally, we state the main result of this section regarding the gradient estimates, $G_i(w) = (X_i, w) - Y_i^t X_i$, for any candidate weight vector $w$, which can be proved by combining the aforementioned results:
Lemma 5.5. Let $\mathcal{D}$ satisfy Assumption 2.2. Given $n = \tilde{O}(d/\eta)$ $(2\eta)$-corrupted samples from $\mathcal{D}$, with probability at least 0.95, there exists for every $w \in \mathbb{R}^d$, a subset $S \subset [n]$ such that $|S| = (1 - 5\eta)n$ satisfying:

$$\|\mathbb{E}_S[G_i(w)] - G^*(w)\| \leq O(\sqrt{\eta}(\|w - w^*\| + \sigma)),$$
$$\|\mathbb{E}_S[G_i(w)G_i(w)^\top]\| \leq O(\|w - w^*\|^2 + \sigma^2),$$

and furthermore, for all $i \in S$, we have $\|G_i(w)\| \leq O\left(\sqrt{\frac{d}{\eta}}\|w - w^*\| + \sigma\sqrt{\frac{d}{\eta}}\right)$.

With this deterministic result in hand we can directly conduct the algorithmic analysis.

References

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time, 2020.

S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In Proceedings of the 2017 Conference on Learning Theory, volume 65 of Proceedings of Machine Learning Research, pages 169–212. PMLR, 07–10 Jul 2017.

K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, and B. Armstrong. Time series regression studies in environmental epidemiology. International journal of epidemiology, 42(4):1187–1195, 2013.

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In Advances in Neural Information Processing Systems 28, pages 721–729, 2015.

K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In Advances in Neural Information Processing Systems 30, pages 2110–2119, 2017.

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

Y. Cheng, I. Diakonikolas, and R. Ge. High-dimensional robust mean estimation in nearly-linear time. In Proceedings of the Thirty-first Annual ACM-SIAM Symposium on Discrete Algorithms, pages 2755–2771. SIAM, 2019.

I. Diakonikolas and D. M Kane. Recent advances in algorithmic high-dimensional robust statistics. arXiv preprint arXiv:1911.05911, 2019.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In Foundations of Computer Science (FOCS), 2016.

I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 999–1008. PMLR, 2017.

I. Diakonikolas, G. Kamath, D. M Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 2683–2702. SIAM, 2018a.
I. Diakonikolas, D. M Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018b. 3

I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1596–1606. PMLR, 09–15 Jun 2019a. 2, 3, 6

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’19, page 2745â€”2754, 2019b. 2, 3

T. E Dielman. *Applied Regression Analysis for Business and Economics*. Duxbury/Thomson Learning Pacific Grove, CA, 2001. 1

Y. Dong, S. Hopkins, and J. Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *Advances in Neural Information Processing Systems 3*, 2019. 2, 3, 5

C. Gao. Robust regression via multivariate regression depth, 2017. 5

C. F. Gauß. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Hamburgi : sumtibus Frid. Perthes et I. H. Besser, 1809, 1809. 1

S. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018. 3

P. J Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101, 1964. 2

P. J Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004. 3

T. Isobe, E. D Feigelson, M. G Akritas, and G. J. Babu. Linear regression in astronomy. *The astrophysical journal*, 364:104–113, 1990. 1

A. Klivans, P. K. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1420–1430. PMLR, 2018. 2, 3

P. K Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018. 3

K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS)*, 2016. 3

G. Lecué and J. Depersin. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019. 2, 12

G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016. 4
In this section, we complete the proofs of Theorems 2.3 and 2.5 given the results established in previous sections.

A.1 Proof of Theorem 2.3

First, we get that the conclusions of Lemma 5.5 holds for the sample points with probability at least 0.95. Now, in light of Lemma 4.3, we only need to ensure that one obtains a good solution to MT in for all iterations from 0 to T. That is we need to ensure that we obtain a good solution to MT in each iteration. Lemma E.6 guarantees a good solution in a single run with probability 0.9. To boost the probability, we will simply run
the Algorithm in Lemma E.6, \(O(\log T)\) times each iteration. We obtain a good solution in at least one run with probability at least \(1 - 1/(1000T)\). To ensure we select a good solution, we compute \(A_{\max}(\mathbb{E}_s[G^T_i G_i])\) for each solution \(s \in \Delta_{100}\) and select the one with the smallest value. From the union bound, the probability that one has a good solution to \(MT\) for all \(T\) rounds is at least 0.999. By taking a union bound over the events described by Eq. (4), Lemma 5.5 and the success of the solver designed for \(MT\), we conclude that Algorithm 1 successfully returns an estimate, \(\hat{w}\), satisfying \(\|\hat{w} - w^*\| \leq O(\sigma \kappa \sqrt{\eta})\) with probability at least 0.9.

\(\Box\)

A.2 Proof of Theorem 2.5

We first analyze the set of samples, \(\tilde{Z}_1\). From Lemma B.1, we get that \(\mathbb{P}(\|X\| \leq O(\sqrt{d})) \geq 0.99\). Therefore, we have via Hoeffding’s inequality:

\[
\mathbb{P}\left(\sum_{(X_i, Y_i) \in \tilde{Z}_1} 1\{\|X_i\| \leq O(\sqrt{d})\} \geq 0.95n_1\right) \geq 1 - \exp\left(-\frac{2n}{25^2}\right) \geq 1 - \frac{1}{100d^2}.
\] (4)

We condition now on the above event. Furthermore, for \(\|X_i\| \leq O(\sqrt{d})\), we have that \((X_i, Y_i) \sim D_S\) from Lemma B.1. Since the adversary only corrupts, \(\eta n_1\) sample points, at most \(\eta n_1\) of the 0.95\(n_1\) points in \(\tilde{Z}_1\) are corrupted. From Lemmas B.1 and F.6, we see that Assumption 2.2 also hold for distributions satisfying Assumption 2.4 conditioned on the event \(\{\|X_i\| \leq O(\sqrt{d})\}\). Therefore, the result of Theorem 2.3 also holds for the set of samples \(\tilde{Z}_1\). Therefore, the estimate \(w^*\) in Algorithm 3 satisfies \(\|w^* - w^*\| \leq O(\sqrt{\eta})\) with probability \(9/10\). Therefore, Assumptions 2.4 and 4.4 hold for the set of samples, \(Z_2\). We now see that the conclusion of theorem holds true conditioned on Assumption 4.5 holding for the set of samples \(Z_2\) which is guaranteed with probability at least 0.99 and us obtaining a \((1 + 3\eta)\)-approximate solution to \(MT\) which happens with probability at least \(9/10\) from Lemma E.6. Via a union bound, this event takes place with probability at least \(2/3\) thus proving the theorem.

\(\Box\)

B Heavy-Tailed Concentration Results

Here we collect the proofs of the relevant concentration arguments we use to show the deterministic conditions used in the gradient estimation step for heavy-tailed data. In what follows, we use \(D\) to denote the distribution over the pair \((X, Y)\). We begin by showing that truncating the data doesn’t significantly affect its covariance or hypercontractivity constant.

Lemma B.1. Consider a distribution, \(D\), satisfying all the conditions stated in Assumption 2.2 except the boundedness assumption. Then, there exists a constant \(C_1\) such that the conditional distribution of \(D\) on \(S = \{\|X\| \leq C_1 \sqrt{d}\}\), \(D_S\) satisfies:

1. \(\mathbb{P}(\|X\| \leq C_1 \sqrt{d}) \geq 0.99\)
2. \(0.99\Sigma \leq \mathbb{E}_{D_S}[XX^\top] \leq 1.01\Sigma\)
3. \(\forall \|u\| = 1, \mathbb{E}_{D_S}[\langle X, u \rangle^4] \leq 1.01\tau \cdot (\mathbb{E}_{D_S}[\langle X, u \rangle^2])^2\)
4. For \((X, Y) \sim D_S\), we have that \(\epsilon\) is independent of \(X\) and has the same distribution as when \((X, Y) \sim D\).
**Proof of Lemma B.1.** The fourth claim follows from the fact that $X$ and $\epsilon$ are independent random variables. Formally, we have for any measurable set $A$:

$$
\Pr \left( \epsilon \in A \mid \|X\| \leq C_1 \sqrt{d} \right) = \frac{\Pr \left( \epsilon \in A \cap \|X\| \leq C_1 \sqrt{d} \right)}{\Pr \left( \|X\| \leq C_1 \sqrt{d} \right)} = \frac{\Pr \left( \epsilon \in A \right) \cdot \Pr \left( \|X\| \leq C_1 \sqrt{d} \right)}{\Pr \left( \|X\| \leq C_1 \sqrt{d} \right)} = \Pr \left( \epsilon \in A \right).
$$

To prove the rest of the lemma, first note that we have $\mathbb{E}[\|X\|^2] = \text{tr} \Sigma = d$. From this, we obtain from Markov’s inequality, for any $C > 0$, $\Pr \left( \|X\| \leq C \sqrt{d} \right) \geq 1 - C^{-2}$. Let $S_C = \{ \|X\| \leq C \cdot \sqrt{d} \}$. For the upper bound in the first claim, we have:

$$
\mathbb{E}_{D_{SC}}[XX^\top] = \frac{1}{\mathbb{P}(D_{SC})} \cdot \mathbb{E}_D[XX^\top 1 \{X \in S_C\}] \leq \frac{1}{\mathbb{P}(D_{SC})} \cdot \mathbb{E}_D[XX^\top] \leq \frac{1}{1 - C^{-2} \cdot \Sigma}.
$$

For the lower bound, using the Cauchy-Schwarz inequality, we have for any vector $u$ of norm $\|u\| = 1$:

$$
\mathbb{E}_D[(X,u)^2 1 \{X \notin S_C\}] \leq (\mathbb{E}_D[(X,u)^4])^{1/2}(\mathbb{E}[1 \{X \notin S_C\}])^{1/2} \leq \frac{\sqrt{\Sigma}}{C} \cdot u^\top \Sigma u.
$$

The lower bound now follows since:

$$
\mathbb{E}_{D_{SC}}[(X,u)^2] \geq \mathbb{E}_D[(X,u)^2 1 \{X \in S_C\}] = \mathbb{E}_D[(X,u)^2] - \mathbb{E}_D[(X,u)^2 1 \{X \notin S_C\}] \geq \left(1 - \frac{\sqrt{\Sigma}}{C}\right) \cdot u^\top \Sigma u.
$$

Finally, for the second claim, we have:

$$
\mathbb{E}_{D_{SC}}[(X,u)^4] \leq \frac{1}{1 - C^{-2}} \cdot \mathbb{E}_D[(X,u)^4] \leq \frac{r}{1 - C^{-2}} (u^\top \Sigma u)^2 \leq \frac{r}{1 - C^{-2}} \left(1 - \frac{\sqrt{\Sigma}}{C}\right)^{-2} (\mathbb{E}_{D_{SC}}[(X,u)^2])^2.
$$

By taking $C$ to be sufficiently large, we have the first three claims of the lemma. \qed

We now provide the proof of the first inequality in the deterministic condition.

**Proof of Lemma 5.1.** The upper bound follows from the observation $\mathbb{E}_S[X_iX_i^\top] = \mathbb{E}_{[n]}[X_iX_i^\top]$. In order to apply the matrix Bernstein inequality to control the sum $\mathbb{E}_{[n]}[W_iW_i^\top] - \mathbb{E}[W_iW_i^\top]$ in spectral norm, we first need the almost sure bound $R = \|X_iX_i^\top\| \leq \|X_i\|^2 = O(\frac{\delta}{n})$. Now, letting $Z = XX^\top$, where $X$ is a copy of an i.i.d. $X_i$, we have that the matrix variance satisfies $\frac{1}{n} \|\mathbb{E}[(Z - \mathbb{E}[Z])^2]\| \leq \frac{1}{n} \|\mathbb{E}[Z^2]\| \leq \frac{1}{n} \|\mathbb{E}[||X||^2XX^\top]\| \leq O(\frac{\delta}{n})$ appealing to Lemmas F.1 and F.2. Finally, an application of the matrix Bernstein inequality [see, e.g., Tropp, 2015] shows, $\frac{1}{n} \sum_{i=1}^n X_iX_i^\top \leq \mathbb{E}[XX^\top] + I \cdot O(\sqrt{\sigma^2 \log(2d/\delta)} + R \log(2d/\delta)) \leq \Sigma + \sqrt{\eta} \cdot I$ by taking $\delta = \frac{1}{2d}e$, and $n \geq c\frac{d \log(4d)}{n}$ for large-enough $c$. Finally, $\mathbb{E}_S[X_iX_i^\top] = \frac{1}{n} \mathbb{E}_n[X_iX_i^\top 1\{i \in S\}] \leq \frac{1}{n} \mathbb{E}_n[X_iX_i^\top]$, and since, $\frac{1}{n} \mathbb{E}_n[X_iX_i^\top] \leq \frac{1}{n} + O(\eta)$ we obtain that $\mathbb{E}_S[X_iX_i^\top] \leq \Sigma + \sqrt{\eta} \cdot I$.

For the lower bound, consider a $\delta$-net over the unit sphere with resolution in the embedded $\ell_2$ distance, $G$. For a fixed $u_j \in G$, define the random variable $Z_i = \langle X_i, u_j \rangle^2$. From L4-L2 hypercontractivity and the normalization of the spectral norm of $\Sigma$ we have that $\text{var}(Z_i) \leq \mathbb{E}[Z_i^2] = O(1)$. Now, consider the $1 - 20\eta$ quantile of the random variable $Z_i$, $q$, defined as $\text{Pr}[Z_i \leq q] = 1 - 20\eta$. We can now once again appeal to L4-L2 hypercontractivity and Markov’s inequality to see $20\eta = \text{Pr}[Z_i \geq q] \leq \frac{\mathbb{E}[Z_i^2]}{q} \leq O(\frac{1}{\sqrt{\eta}})$, and $q = O(\frac{1}{\sqrt{\eta}})$. Similarly using the Cauchy-Schwarz inequality, $\mathbb{E}[Z_i 1 \{Z_i \leq q\}] \geq \mathbb{E}[Z_i] - \mathbb{E}[Z_i 1 \{Z_i > q\}] \geq \mathbb{E}[Z_i] - \sqrt{\mathbb{E}[Z_i^2]\text{Pr}[Z_i > q]} \geq \mathbb{E}[Z_i] - O(\sqrt{\eta})$. Since the random variable $1 \{Z_i > q\}$ is almost surely
bounded by 1 and has variance bounded by $20\eta$, applying the Bernstein inequality, we can see that that with probability at least $1 - \exp(-10n\eta)$, at most $(1 - \theta n)$ points will be in the range $[0, q]$. We now apply the Bernstein inequality to the sum $E_{\eta}[Z_1^\dagger\{Z_1 \leq q\}] - E[Z_1\{Z_1 \leq q\}]$ to control its lower tail; note that each element in the sum satisfies $|Z_1\{Z_1 \leq q\} - E[Z_1\{Z_1 \leq q\}]| \leq 2q \leq O\left(\frac{1}{\sqrt{n}}\right)$ and similarly that $\text{var}(Z_1\{Z_1 \leq q\}) \leq O(1)$. An application of the Bernstein inequality shows with probability at least $1 - \exp(-n/10)$, that $\mathbb{E}_{\eta}[Z_1^\dagger\{Z_1 \leq q\}] \geq E[Z_1\{Z_1 \leq q\}] - O(\sqrt{n\eta}) \geq E[Z_1] - O(\sqrt{n\eta})$. Renormalizing the sum by the factor $\frac{n}{|S|} = 1 + O(\eta)$, then shows $\frac{1}{|S|} \sum_{i=1}^n Z_1\{Z_1 \leq q\} \geq E[Z_1] - O(\sqrt{n\eta})$ since $E[Z_1] \leq 1$. Now applying a union bound over the previous two events we obtain the lower bound for fixed $u_j - \delta$ since the at least $10\eta$ points lying in the interval $(q, \infty)$ will be precisely the points which contribute the maximal values to the sum $\frac{1}{|S|} \sum_{i=1}^n Z_i - \delta$ so with probability at least $1 - 2\exp(-n\eta/10)$ for fixed $u_j$, the truncated sum contains the smallest $(1 - 10\eta)n$ of the $Z_i$, which is smaller then any arbitrary set of $(1 - 10\eta)n$ of the $Z_i$. Applying another union bound over the entire $\delta$-net shows that the state holds uniformly for any set of $(1 - 10\eta)n$ samples $S$, and any $u_j \in G$ that $\mathbb{E}_S[Z_i] \geq u_j^\top \Sigma u_j - O(\sqrt{n\eta})$ with probability at least $1 - 2(\frac{3}{\delta})^d \exp(-n/10\eta)$. To lift this to all $v \in S^{d-1}$, define $Q = \mathbb{E}_S[X_i^\top X_i^\top]$ and note that each element $v = u_j + \Delta$ for some $u_j \in G$ and $\Delta$ such that $||\Delta|| \leq \delta$. Thus,

$$\langle v, Qv \rangle = \langle u_j, Qu_j \rangle + 2\langle v, Q\Delta \rangle + \langle \Delta, Q\Delta \rangle \geq \langle u_j, Qu_j \rangle - O(\delta d),$$

uniformly for all $\Delta$ where we have used the crude bound that $||Q|| \leq O(d)$ since the vectors $X_i$ are truncated at $O(\sqrt{n\eta})$. Choosing $\delta = \frac{c}{\sqrt{n}}$ for sufficiently small $c$, then gives the final conclusion that $\frac{1}{|S|} \sum_{i \in S} \langle v, Qv \rangle \geq \sqrt{\delta} - O(\sqrt{n\eta})$ for all $v \in S^{d-1}$ with probability at least $1 - 2(c^2 d^2 / n\eta)^d \exp(-n/10\eta)$. Choosing $n \geq c_2/\eta d \log(2d / \sqrt{n\eta})$ for sufficiently large $c_2$ ensures this event holds with probability at least $1 - \frac{1}{2d^2}$.

We now prove Lemma 5.2 which exchanges the order of the min and max in the SDP used in the gradient estimation step.

**Proof of Lemma 5.2.** First, let $T = \{i : \langle X_i, u \rangle \leq 40 / \sqrt{n}\}$. We now consider the following min-max game:

$$\min_{\alpha \in \Delta_{\eta/25}} \max_{M \geq 0, \text{tr} M = 1} \left\langle E_{\alpha}[\langle X_i, u \rangle^2 X_i X_i^\top], M \right\rangle = \max_{M \geq 0, \text{tr} M = 1} \min_{\alpha \in \Delta_{\eta/25}} \left\langle E_{\alpha}[\langle X_i, u \rangle^2 X_i X_i^\top], M \right\rangle,$$

where the min and the max can be exchanged as a consequence of von Neumann’s min-max theorem. Assume now that the conclusion of Lemma 5.4 holds. Then, we have that the right hand side of the above display is upper bounded by $C_6$ and there exists $\alpha^* \in \Delta_{\eta/25}$ such that:

$$E_{\alpha^*}[\langle X_i, u \rangle^2 X_i X_i^\top] \leq C_6.$$

Now, let $S = \{i \in T : \alpha_i \geq 1/(2n)\}$. Note that $S$ being a subset of $T$ automatically proves the second claim of the lemma. We have from the fact that $|T| \geq (1 - \eta/25)n$:

$$1 - \frac{\eta}{25 - \eta} = \sum_{i \in S} \alpha_i + \sum_{i \not\in S} \alpha_i \leq \frac{|T \setminus S|}{2n} + \frac{|T| - |T \setminus S|}{(1 - \eta/25)n} \implies |T \setminus S| \leq (\eta n)/6,$$

where the inequality follows when $\eta$ is sufficiently small and the fact that $|T| \leq n$. Finally, the statement of the lemma follows with $C_5 = 3C_6$:

$$\mathbb{E}_S[\langle X_i, u \rangle^2 X_i X_i^\top] \leq 3 \mathbb{E}_{\alpha^*}[\langle X_i, u \rangle^2 X_i X_i^\top] \leq 3C_6.$$

□
We now provide the proof of Lemma 5.3 which shows there exist a set of weights which can exclude problematic $X_i$ in the normalized sums we consider.

**Proof of Lemma 5.3.** We first grid the unit spheres $G_1 = \|u\| \leq 1$ and $G_2 = \|v\| = 1$ with grids of resolution $1/(100C^4_d d^2)$. That is, for all $\|u\| \leq 1$, there exists $u' \in G_1$ such that $\|u - u'\| \leq 1/(100C^4_d d^2)$ and similarly for $v$. Now, let $q$ be defined as follows:

$$q = \min \{ r : \forall \|u\| = 1, \mathbb{P}(\|X_i, u\| \leq r) \geq 1 - \eta/400 \}.$$ 

We know that $q \leq 20/\sqrt{\eta}$ by Markov’s inequality applied to $\langle X_i, u \rangle^2$. Now, for fixed $u$ and $v$, consider the random variables $Q_i = 1 \{ |\langle X_i, u \rangle| < 1000/\sqrt{\eta} \} \wedge \langle X_i, v \rangle \leq 1000/\sqrt{\eta}$. By the union bound, we have $\mathbb{P}(Q_i = 1) \geq 1 - \eta/200$. Furthermore, consider the random variables $W_i = \langle X_i, u \rangle^2 \langle X_i, v \rangle^2 Q_i$. We see that $W_i$ is bounded by $1000^2/\eta$ and furthermore has variance bounded as follows:

$$\mathbb{E}[W_i^2] \leq \frac{1000^2}{\eta} \mathbb{E}[\langle X_i, v \rangle^4] \leq \frac{\tau 1000^2}{\eta} (v^T \Sigma v)^2 \leq \frac{\tau 1000^2}{\eta}.$$ 

Furthermore, its mean is bounded by Cauchy Schwarz and L4-L2 hypercontractivity as:

$$E[W_i] \leq E[\langle X_i, u \rangle^2 \langle X_i, v \rangle^2] \leq \tau \cdot (u^T \Sigma u)(v^T \Sigma v) \leq \tau.$$ 

Therefore, applying Bernstein’s inequality to the random variable $W = n^{-1} \sum_{i=1}^n W_i$, we get:

$$\mathbb{P}(W \geq 1 + \tau) \leq \exp \left( - \frac{\eta \tau^2}{2 \cdot 1000^2 (\tau + 1)} \right).$$

(5) 

Finally, consider the random variable $K_i = 1 \{ |\langle X_i, u \rangle| \geq q \lor |\langle X_i, v \rangle| \geq q \}$. We see that $K_i$ has variance and expectation bounded by $\eta/200$ and is bounded by 1. Therefore, applying Bernstein’s inequality to $K = n^{-1} \sum_{i=1}^n K_i$, we get that:

$$\mathbb{P} \left( K \geq \frac{\eta}{100} \right) \leq \exp \left( - \frac{\eta^2 n^2}{200 (\eta n/200 + \eta n/200)} \right) = \exp \left( - \frac{\eta n}{800} \right).$$

(6) 

From standard packing/covering number bounds [see, e.g., Vershynin, 2012], we see that the number of elements in each grid is at most $(C d)^{2d}$ for some constant $C$. Therefore, the total number of pairs of elements, $(u, v)$, from each grid is at most $(C d)^{4d}$. Therefore, taking $n = O(d/\eta)$, we have that for all $u \in G_1$ and $v \in G_2$ the events described in Eqs. (5) and (6) hold with probability at least $1 - 1/d$. Now, consider any $\|u\| \leq 1$ and $\|v\| \leq 1$. Let $u' \in G_1$ and $v' \in G_2$ be such that $\|u' - u\| \leq 1/(100C^4_d d^2)$ and $\|v' - v\| \leq 1/(100C^4_d d^2)$. Now, we have for any $X_i$:

$$\langle X_i, u \rangle^2 \langle X_i, v \rangle^2 - \langle X_i, u' \rangle^2 \langle X_i, v' \rangle^2 = (\langle X_i, u \rangle^2 \langle X_i, v \rangle^2 - \langle X_i, u \rangle \langle X_i, u \rangle \langle X_i, v \rangle + \langle X_i, u \rangle \langle X_i, v \rangle \langle X_i, v \rangle - \langle X_i, u \rangle^2 \langle X_i, v \rangle^2) + (\langle X_i, u \rangle^2 \langle X_i, v \rangle^2 - \langle X_i, u' \rangle^2 \langle X_i, v' \rangle^2) + (\langle X_i, u \rangle^2 \langle X_i, v \rangle^2 - \langle X_i, u' \rangle^2 \langle X_i, v' \rangle^2).$$

Now, each of the four terms is bounded by $\max(\|u' - u\|, \|v' - v\|) \|X_i\|^4$ using the Cauchy Schwarz inequality. From our bounds on $\|u' - u\|$, $\|v' - v\|$ and $\|X_i\|$, we obtain for all $i$:

$$\langle X_i, u \rangle^2 \langle X_i, v \rangle^2 - \langle X_i, u' \rangle^2 \langle X_i, v' \rangle^2 \leq 1.$$ 

(7)
We also have, using Cauchy Schwarz inequality:

\[ |\langle X_i, u \rangle - \langle X_i, u' \rangle| \leq \|u - u'\| \|X_i\| \leq 1, \tag{8} \]

with a similar bound holding for \( v \). From Eqs. (7) and (8), we have:

\[
n^{-1} \sum_{i=1}^{n} (X_i, u)^2 (X_i, v)^2 1 \{ |\langle X_i, u \rangle| \leq 800/\sqrt{\eta} \land |\langle X_i, v \rangle| \leq 800/\sqrt{\eta} \} \leq 1 + n^{-1} \sum_{i=1}^{n} (X_i, u')^2 (X_i, v')^2 1 \{ |\langle X_i, u \rangle| \leq 800/\sqrt{\eta} \land |\langle X_i, v \rangle| \leq 800/\sqrt{\eta} \} \leq 1 + n^{-1} \sum_{i=1}^{n} (X_i, u')^2 (X_i, v')^2 1 \{ |\langle X_i, u' \rangle| \leq 1000/\sqrt{\eta} \land |\langle X_i, v' \rangle| \leq 1000/\sqrt{\eta} \} \leq \tau + 2.
\]

For the second statement of the lemma, we have from Eq. (8):

\[
\sum_{i=1}^{n} 1 \{ |\langle X_i, u \rangle| \leq 40/\sqrt{\eta} \land |\langle X_i, v \rangle| \leq 40/\sqrt{\eta} \} \geq \sum_{i=1}^{n} 1 \{ |\langle X_i, u' \rangle| \geq \eta \land |\langle X_i, v' \rangle| \geq \eta \} \geq n(1 - \eta/100).
\]

This concludes the proof of the lemma. \(\square\)

We now present the proof of Lemma 5.4 which allows us to lift the previous argument which was restricted to rank-one \( M \) to general \( M \).

Proof of Lemma 5.4. We first condition on the event which holds in the conclusion of Lemma 5.3 and consider a vector \( u \) of norm \( \|u\| = 1 \) and a trace-1 psd matrix \( M \). Let \( g \sim N(0, M) \) be a centered gaussian random vector with covariance matrix \( M \). We have by an application of the Borell-TIS inequality [see, e.g., Ledoux and Talagrand, 1991] that with probability at least \( 1 - \exp(-9/2) \geq 0.95, \|g\| \leq 4 \). Furthermore, for any \( i \), we have that \( \langle X_i, g \rangle \) is gaussian with mean 0 and variance \( \langle X_i, X_i^\top, M \rangle \). Therefore, we get:

\[
\mathbb{P}\left( |\langle g, X_i \rangle| \geq 0.5 \sqrt{\langle X_i, X_i^\top, M \rangle} \right) \geq 1 - 1/\sqrt{2\pi} \geq 0.6.
\]

So, for any \( i \), we have:

\[
\mathbb{P}\left( |\langle g, X_i \rangle| \geq 0.5 \sqrt{\langle X_i, X_i^\top, M \rangle} \land \|g\| \leq 4 \right) \geq 0.5.
\]

Let \( I \) denote the indices such that \( \sqrt{\langle X_i, X_i^\top, M \rangle} \geq 400/\sqrt{\eta} \). Then, we have:

\[
\mathbb{E}\left[ \sum_{i \in I} 1 \{ |\langle g, X_i \rangle| \geq 200/\sqrt{\eta} \land \|g\| \leq 4 \} \right] = \mathbb{E}\left[ \sum_{i \in I} 1 \{ |\langle g, X_i \rangle| \geq 50/\sqrt{\eta} \land \|g\| \leq 4 \} \right] \geq \frac{1}{2} |I|.
\]

From the second conclusion of Lemma 5.3, each term inside the expectation is bounded by \( \frac{\eta}{200} \cdot n \). Therefore, we have \( |I| \leq \frac{4n}{\eta} \). This establishes the second claim of the lemma as there are at most \( \frac{n}{100} \) points, \( X_i \), with \( |\langle X_i, u \rangle| \geq 40/\eta \).

Now, for \( j \in [n] \setminus I \), we have:

\[
\mathbb{E}[ (X_j, g)^2 1 \{ \|g\| \leq 4 \land |\langle X_j, g \rangle| \leq 800/\sqrt{\eta} \}] \geq \mathbb{E}[ (X_j, g)^2 1 \{ \|g\| \leq 4 \land |\langle X_j, g \rangle| \leq 800/\sqrt{\eta} \}]^2
\]
\[
\langle X_j, X_j^\top, M \rangle \cdot \frac{1}{2\pi} \cdot \left( \int_{-1}^{1} |x| \exp \left( -\frac{x^2}{2} \right) \, dx \right)^2 \geq \frac{1}{4\pi} \cdot \langle X_j X_j^\top, M \rangle,
\]

this is because the event \{\|g\| \leq 4 \wedge |\langle X_j, g \rangle| \leq 800/\sqrt{n}\} happens with probability at least 0.9 and it can be seen via a simple monotonicity argument that the value of \( \mathbb{E}[Z^2 1(A)] \) for \( Z \sim \mathcal{N}(0, 1) \) is minimized when \( A \) is chosen as an interval around the origin with probability mass \( \mathbb{P}(A) \).

From the previous display, we finally get:

\[
16C_4 \geq 16 \cdot \mathbb{E} \left[ \sum_{j \notin I} \left( \frac{gg^\top}{16} \cdot X_j X_j^\top \right) 1 \{ \|g\| \leq 4 \wedge |\langle X_j, g \rangle| \leq 800/\sqrt{n} \} \right] \geq \frac{1}{4\pi} \sum_{j \notin I} \langle X_j X_j^\top, M \rangle.
\]

By rearranging the above display, we get the first claim of the lemma with \( C_5 = 64\pi C_4 \).

Here we state and prove the result needed to show there is a good set of weights to remove errors from terms of the form \( \epsilon_i X_i \) in the gradient term.

**Lemma B.2.** Let \( D \) satisfy Assumption 2.2. Then, for \( n = \tilde{O}(d/\eta) \), we have with probability 99/100, there exists a set \( S \subset [n] \) with \( |S| \geq 1 - \eta/100 \) such that for any set \( T \subset S \) with \( |T| \geq (1 - 10\eta) \), we have:

\[
\|T\|^{-1} \sum_{i \in T} \epsilon_i X_i \| \leq 15\sqrt{\eta} \sigma \text{ and } |T|^{-1} \sum_{i \in T} \epsilon_i^2 X_i X_i^\top \leq 5 \sigma^2 I.
\]

And furthermore, we have \( \|\epsilon_i X_i\| \leq O(\sigma \sqrt{d/\eta}) \) for all \( i \in S \).

**Proof of Lemma B.2.** We first note that \( Z_i = \epsilon_i X_i \) is a random vector with zero mean and variance \( \sigma^2 \Sigma \). Therefore, we have that \( \mathbb{P}(\|Z_i\| \leq 10\sigma \sqrt{d/\eta}) \geq 1 - \eta/200 \) by Markov’s Inequality applied to \( \|Z_i\|^2 \). Now, our set \( S \) will be defined as follows \( S = \{i : \|Z_i\| \leq 10\sigma \sqrt{d/\eta} \} \). Let \( W_i = 1\{\|Z_i\| \leq 10\sigma \sqrt{d/\eta} \} \). We have \( P(W_i = 1) \geq 1 - \eta/100 \) and \( W_i \) has variance bounded by \( \eta/100 \). Therefore, we get by Bernstein’s inequality:

\[
P\left( \sum_{i=1}^{n} W_i - 1 - \eta/50 \right) \leq \exp \left( -\frac{1}{100^2} \cdot \frac{\eta^2 n^2}{2(n \cdot \eta/100 + n \cdot \eta/100)} \right) = \exp \left( -\frac{\eta n}{400} \right).
\]

Therefore, we have with probability at least \( 1 - 1/e^d \). Let \( A = \{Z : \|Z\| \leq 10\sigma \sqrt{d/\eta} \} \). Now, conditioned on \( Z_i \in A \), the conditional mean and second moment matrix of \( Z_i \) are defined as follows for \( Z_k \sim D \):

\[
\mu_A = \frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[Z_k 1 \{Z_k \in A\}] \text{ and } \frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[Z_k Z_k^\top 1 \{Z_k \in A\}],
\]

and furthermore, all the vectors in \( S \) are independent of each other. We will now bound the mean term in the above display. Take \( \|u\| = 1 \):

\[
\frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[\langle Z_k, u \rangle 1 \{Z_k \in A\}] = \frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[\langle Z_k, u \rangle (1 - 1 \{Z_k \notin A\})]
\]

\[
= \frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[\langle Z_k, u \rangle 1 \{Z_k \notin A\}]
\]

\[
\leq \frac{1}{\mathbb{P}(Z_k \in A)} (\mathbb{E}[\langle Z_k, u \rangle^2])^{1/2} (\mathbb{E}[1 \{Z_k \notin A\}])^{1/2} \leq 2\sqrt{\eta} \sigma.
\]
where the first inequality follows from Cauchy-Schwarz. Therefore, the mean of $Z_i$ for $Z_i \in S$ is bounded by $2\sqrt{n}$. For the variance, we have:

$$\frac{1}{\mathbb{P}(Z_k \in A)} \mathbb{E}[Z_k Z_k^\top 1\{Z_k \in A\}] \leq 2 \cdot \mathbb{E}[Z_k Z_k^\top] = 2\sigma^2 \Sigma.$$

Therefore, we have with probability at least 0.999 that $|S|^{-1} \sum_{i \in S} Z_i \leq 4\sigma \sqrt{n}$. Furthermore, we have by matrix Bernstein [Tropp 2015] that:

$$\mathbb{P}\left( \frac{1}{|S|} \sum_{i \in S} Z_i Z_i^\top \leq 4\sigma^2 \cdot I \right) \geq 1 - 2d \exp\left( - \frac{d n^2}{16(4n + 40nd/\eta)} \right) \geq 1 - \frac{1}{d^2},$$

where we have used $(\mathbb{P}(Z_k \in A))^{-1}\mathbb{E}[\|Z_k\|^2 Z_k Z_k^\top] \leq 2\mathbb{E}[\|Z_k\|^2 Z_k Z_k^\top]$ and Lemma F.2 to bound the matrix variance term. Therefore, we have with probability at least 0.99 that there exists a set $S \subset [n]$ with $|S| \geq (1 - \eta/100)n$ such that:

$$\mathbb{E}_S[Z_i] \leq 4\sigma \sqrt{n} \text{ and } \mathbb{E}_S[Z_i Z_i^\top] \leq 4\sigma^2 I.$$

Finally, we have for any $T \subset S$ with $|T| \geq (1 - 10\eta)n$ and any $\|u\| = 1$:

$$\mathbb{E}_T[\langle u, Z_i \rangle] = \frac{1}{\mathbb{P}_S(Z_i \in T)} \mathbb{E}_S[\langle u, Z_i \rangle(1 - 1\{i \notin T\})] \leq 5\sigma \sqrt{n} + 1.01(\mathbb{E}_S[\|u, Z_i\|^2])^{1/2}(\mathbb{E}_S[1\{i \notin T\}])^{1/2} \leq 15\sigma \sqrt{n}.$$

The upper bound on the second moment follows from the fact that $\mathbb{E}_T[Z_i Z_i^\top] \leq \frac{1}{\mathbb{P}_S(Z_i \in T)} \mathbb{E}_S[Z_i Z_i^\top]$.

We can now assemble the previous results to provide the proof of the final lemma which establishes the deterministic conditions:

**Proof of Lemma 5.5.** Under the assumptions, the conclusions of Lemmas 5.1, 5.2 and B.2 all hold with probability 0.99. Let $u = \frac{w-w^*}{\|w-w^*\|}$ and label the set guaranteed to exist by Lemma B.2 be $S_1$ and the one by Lemma B.2 by $S_2$. And consider the intersection of the two sets $S_3 = S_1 \cap S_2$. By the union bound, $|S_3| \geq 1 - \eta/5$. Since, at most $2\eta n$ of the elements of $S_3$ are corrupted, there exists a set $S \subset S_3$ of size at least $(1 - 5\eta)n$ such that none of their elements have been corrupted. Now, we have for $i \in S$:

$$\|G_i\| = \|X_i \langle X_i, w - w^* \rangle - \epsilon_i X_i\| \leq \|X_i \langle X_i, w - w^* \rangle\| + \|\epsilon_i X_i\| \leq O\left(\frac{\sqrt{d}}{\eta} \|w - w^*\| + \sigma \sqrt{\frac{d}{\eta}}\right).$$

This proves the third claim of the lemma. For the first claim, we have:

$$||S|^{-1} \sum_{i \in S} G_i(w) - \Sigma(w - w^*)|| \leq |||S|^{-1} X_i X_i^\top - \Sigma||\|w - w^*\| + |||S|^{-1} (\epsilon_i X_i)||$$

$$\leq |||S|^{-1} X_i X_i^\top - \Sigma||\|w - w^*\| + O(\sqrt{\eta} \sigma) \leq O(\sqrt{\eta}||w - w^*\| + \sigma)).$$

Finally, for the second claim of the lemma, we have:

$$|S|^{-1} \sum_{i \in S} G_i(w)G_i(w)^\top \leq 2|S|^{-1} \sum_{i \in S} (X_i, w - w^*)^2 X_i X_i^\top + \epsilon_i^2 X_i X_i^\top \leq O(||w - w^*||^2 + \sigma^2) \cdot I.$$

With this result in hand we can establish the deterministic conditions necessary for our algorithms hold with the requisite probability.
C Sub-Gaussian Concentration

In this section we prove concentration results pertaining to the case where the random variables, \( X \) and \( \epsilon \) are sub-Gaussian instead of the much milder L4-L2 assumptions we had previously considered. This allows us to obtain much tighter bounds on our recovery error than those obtained in Section 5.

From Theorem 2.3, we can assume that we have an initial estimate \( w \) such that \( (w - w^\star) \leq O(\sqrt{\eta}) \). Therefore, by subtracting out \( \langle X_i, w \rangle \) from all the data points, our problem reduces to a setting where we additionally have \( \|w^\star\| \leq O(\sqrt{\eta}) \). We make this formal in the following assumption:

**Assumption C.1.** We assume that there exists a constant \( \nu \), such that \( \|w^\star\| \leq \nu \sigma \sqrt{\eta} \).

Note, that under Assumption 2.4, we have \( \mathbb{E}[\|X\|] \leq \left( \mathbb{E}[\|X\|^2] \right)^{1/2} = \sqrt{d} \). Throughout the analysis we refer at several points to the following events:

\[
\mathcal{H}_e = \{ |\epsilon| \leq 6\phi \sqrt{\log 1/\eta} \}
\]
\[
\mathcal{H}_l = \{ \|X\| \leq 7\psi \sqrt{d} + 24\psi \sqrt{\log 1/\eta} \}
\]
\[
\mathcal{H}_o = \{ |\langle X, w - w^\star \rangle| \leq 6\psi \|w - w^\star\| \sqrt{\log 1/\eta} \}.
\]

Note, that from Lemma C.3 the event, \( \mathcal{H} = \mathcal{H}_e \cap \mathcal{H}_l \cap \mathcal{H}_o \) occurs with probability at least \( 1 - \eta^{16} \). We start by proving a lemma pertaining to the noise terms arising in the gradient.

**Lemma C.2.** Let \( D \) satisfy Assumption 2.4. Then for \( n = \tilde{O}(d/\eta^2) \) samples from \( D \), we have with probability at least \( 1 - 3/(10d^2) \), for any set \( S \subset [n] \) with \( |S| \geq (1 - 10\eta)n \), we have:

\[
\left\| \sum_{i \in S} Z_i/n \right\| \leq O(\sigma \eta \log 1/\eta) \quad \text{and} \quad \left\| \sum_{i \in S} Z_i z_i^\top/n - \sigma^2 I \right\| \leq O(\sigma^2 \eta \log^2 1/\eta),
\]

where \( z_i = \epsilon_i X_i 1 \{(\epsilon_i, X_i) \in \mathcal{H} \} \).

**Proof of Lemma C.2.** Consider the random matrix \( M = n^{-1} \sum_{i=1}^n M_i \) where the matrices \( M_i \) are defined by \( M_i = \epsilon_i^2 X_i x_i^\top 1 \{(X_i, \epsilon_i) \in \mathcal{H} \} \). From the definition of the event \( \mathcal{H} \), we can bound each of them as \( M_i \leq C \phi^2 \psi^2 \log 1/\eta(d + \log 1/\eta) \). Furthermore, we have \( \mathbb{E}[M_i] \leq \mathbb{E}[\epsilon_i^2 X_i x_i^\top] = \sigma^2 \cdot I \). For a lower bound, we have for any unit vector \( u \), by Cauchy-Schwarz inequality and standard bounds on the moments of sub-Gaussian distributions:

\[
\mathbb{E}[u^\top M_i u] = \mathbb{E}[\epsilon_i^2 \langle X_i, u \rangle^2 (1 - 1 \{X_i \notin \mathcal{H} \})] \geq \sigma^2 - (\mathbb{E}[\epsilon_i^4] \mathbb{E}[\langle X_i, u \rangle^4])^{1/2} (1 - \mathbb{P}(\mathcal{H}))^{1/2} \geq \sigma^2 (1 - \eta^8).
\]

Therefore, applying the matrix Bernstein inequality [see, e.g., Tropp, 2015] leads to:

\[
\mathbb{P} \left( \|M - \sigma^2 \cdot I\| \geq \sigma^2 \eta \right) \leq 2d \exp \left( -\frac{\eta^2 n^2}{4C'(nd + \eta \log 1/\eta(d + \log 1/\eta))} \right) \leq 1 - \frac{1}{10d^2},
\]

where we have used the fact that sub-Gaussian random variables are L4-L2 hypercontractive (Lemma F.6), Lemma F.2, Assumption 2.4 and our setting of \( n \).

For the lower bound, we begin by gridding the unit sphere with a grid of resolution \( \left( \frac{\eta}{2} \right)^\delta \). From standard results, there exists a grid, \( G \) of size \( \left( C^4 d \frac{n}{\eta} \right)^{8d} \) for some absolute constant \( C^4 \). Now, for a specific \( u \in G \), consider the random variable \( W_i = \epsilon_i^2 \langle X_i, u \rangle^2 1 \{\mathcal{H} \cap \mathcal{H}_{ui} \} \) where \( \mathcal{H}_{ui} = \{ |\langle X_i, u \rangle| \leq 6\psi \sqrt{\log 1/\eta} \} \). We see
that \( P(\mathcal{H}_u) \geq 1 - \eta^{16} \). Therefore, we have by a similar argument as before that \( P(\mathcal{H} \cap \mathcal{H}_u) \geq 1 - 2\eta^8 \).

Furthermore, note that \( W_i \) is upper bounded by \( 36^2\psi^2\phi^2 \log^2 1/\eta \). Therefore, applying Hoeffding’s inequality to \( W = n^{-1} \sum_{i=1}^n W_i \), we get:

\[
P(W \leq \sigma^2(1 - \eta)) \leq \exp \left( -\frac{mn^2}{(36^2\psi^2\phi^2 \log^2 1/\eta)^2} \right).
\]

By taking a union bound over \( \mathcal{G} \), we see that the above concentration holds for all \( u \in \mathcal{G} \) with probability at least \( 1 - 1/(10d^2) \). Now, for a non-grid element \( v \), select \( v' \in \mathcal{G} \) such that \( \|v - v'\| \leq (\frac{\eta}{d})^8 \). Note, that for any \( (X_i, \epsilon_i) \in \mathcal{H} \), we have that \( |\langle X_i, v \rangle - \langle X_i, v' \rangle| \leq \epsilon^4 \) and \( |\epsilon_i^2 \langle X_i, v \rangle^2 - \epsilon_i^2 \langle X_i, v' \rangle^2| \leq \epsilon^4 \). Therefore, we conclude that for all \( \|u\| = 1 \), we have:

\[
\sum_{i=1}^n \epsilon_i^2 \langle X_i, u \rangle^2 1 \{\langle X_i, \epsilon_i \rangle \in \mathcal{H} \wedge |\langle X_i, u \rangle| \leq 7\psi \sqrt{\log 1/\eta}\} \geq 1 - 2\eta.
\]

Therefore, the removal of \( 10\eta n \) points at most distorts \( W \) by a factor of \( 10\eta \cdot 36^2\phi^2\psi^2 \log^2 1/\eta \), thus proving the second claim of the lemma.

For the first claim, consider the random vector \( Z_i = \epsilon_i X_j 1 \{\langle X_i, \epsilon_i \rangle \in \mathcal{H} \} \). To bound the mean of \( Z_i \), we consider as before a unit vector \( u \):

\[
\mathbb{E}[\langle u, Z_i \rangle 1 \{\langle X_i, \epsilon_i \rangle \in \mathcal{H} \}] \leq (\mathbb{E}[\langle u, Z_i \rangle ]^2)^{1/2} (P(\mathcal{H}))^{1/2} \leq \eta^8 \sigma.
\]

Since, \( u \) is arbitrary, we have that the mean of the norm of \( Z_i \) is bounded by \( \eta^8 \sigma \). Also, note from the definition of \( Z_i \) that \( Z_i \) is sub-Gaussian with sub-Gaussian parameter at most \( 6\phi\psi \sqrt{\log 1/\eta} \). Therefore, we have with probability at most \( 1 - 1/(10d^2) \) from Lemma F.4 that \( \|n^{-1} \sum_{i=1}^n Z_i\| \leq \phi\psi \eta \). Now, we have for any \( T \subset [n] \) with \( |T| \geq (1 - 10\eta)n \) and any \( \|u\| = 1 \):

\[
\langle u, \mathbb{E}_{[n]}[Z_i] - \mathbb{E}_{[n]}[Z_i 1 \{i \notin T\}] \rangle = \mathbb{E}_{[n]}[\langle u, Z_i \rangle 1 \{i \notin T\}]
\leq (\mathbb{E}_{[n]}[\langle u, Z_i \rangle^2 1 \{i \notin T\}])^{1/2} (\mathbb{E}_{[n]}[1 1 \{i \notin T\}])^{1/2}
\leq (\sigma^2 \eta + 10\eta \cdot 36^2\phi^2\psi^2 \log^2 1/\eta)^{1/2} \sqrt{10\eta}
\leq 20\phi\psi \eta \log 1/\eta,
\]

where the first inequality follows from Cauchy-Schwarz and the second inequality follows from the second claim of the lemma. This concludes the proof of the lemma. \( \square \)

The following lemma is the analogue Lemma 5.1 for the sub-Gaussian case. It’s proof follows similarly to Lemma 5.1.

**Lemma C.3.** Let \( \{X_i\}_{i=1}^n \) satisfy Assumption 2.4. Then, there exists a universal constant \( c \) such that if \( n = \tilde{O}(d/\eta^2) \), with probability at least \( 1 - 1/(10d^2) \) for any set of \( (1 - 10\eta)n \) samples, \( S \), we have:

\[
(1 - O(\eta \log 1/\eta))\Sigma \leq \mathbb{E}_{S}[X_i X_i^\top 1 \{\langle X_i, \epsilon_i \rangle \in \mathcal{H} \}] \leq (1 + O(\eta))\Sigma.
\]

**Proof of Lemma C.3.** For the upper bound, consider the random matrix, \( Z_i = X_i X_i^\top 1 \{\langle X_i, \epsilon_i \rangle \in \mathcal{H} \} \). We first notice that \( \|Z_i\| = \|X_i\|^2 \leq O(d + \log 1/\eta) \) from the definition of \( \mathcal{H} \). Furthermore, we have from Lemma F.2 and Lemma F.6, that \( \|\mathbb{E}[Z_i] - \mathbb{E}[Z_i] 1 \| \leq \|\mathbb{E}[Z_i] 2 \| \leq O(d) \). Therefore, we may apply the matrix Bernstein inequality to \( Z = \mathbb{E}_{[n]}[Z_i] \) to obtain:

\[
P(\|Z - \mathbb{E}[Z]\| \geq \eta) \leq d \exp \left( -\frac{n^2\eta^2}{C(d n + (d + \log 1/\eta)n)} \right) \leq \frac{1}{100d^2}.
\]
for our value of $n$. To bound $\mathbb{E}[Z_i]$, we have for all $u$:

$$|1 - \mathbb{E}[u^T Z_i u]| = |\mathbb{E}[\langle X_i, u \rangle^2 1 \{ (X_i, \epsilon_i) \notin \mathcal{H} \}]| \leq (\mathbb{E}[\langle X_i, u \rangle^4])^{1/2} (\mathbb{P}(\langle X_i, \epsilon_i \rangle \notin \mathcal{H}))^{1/2} \leq \eta^6,$$

from Lemma F.6 and the definition of $\mathcal{H}$. The upper bound now follows from the fact that $\mathbb{E}_S[Z_i] \leq (1 - 10\eta)^{-1} \mathbb{E}_{[n]}[Z_i]$.

For the lower bound, we begin by constructing a $\delta$-net of the unit sphere of resolution $\left(\frac{d}{\eta}\right)^8$. As in previous lemmas, there exists a grid $\mathcal{G}$ of size $\left(\frac{C^2 d}{\eta}\right)^{8d}$. Now, for a particular $u \in \mathcal{G}$, consider the random variable $W_i = \langle X_i, u \rangle^2 1 \{ \mathcal{H} \cap \mathcal{H}_u \}$ where $\mathcal{H}_u = 1 \{ |\langle X_i, u \rangle| \leq 6\psi \sqrt{\log 1/\eta} \}$. We have from Lemma F.5 that:

$$|1 - \mathbb{E}[W_i]| \leq |\mathbb{E}[\langle X_i, u \rangle^2 1 \{ (X_i, \epsilon_i) \notin \mathcal{H} \cap \mathcal{H}_u \}]| \leq (\mathbb{E}[\langle X_i, u \rangle^4])^{1/2} (\mathbb{P}(\langle X_i, \epsilon_i \rangle \notin \mathcal{H} \cap \mathcal{H}_u))^{1/2} \leq \eta^4,$$

from Lemma F.6 and the definition of $\mathcal{H}$ and $\mathcal{H}_u$. Now, applying Hoeffding’s inequality to $W = \mathbb{E}_{[n]}[W_i]$, we get:

$$\mathbb{P}(\|W - \mathbb{E}[W]\| \geq \eta) \leq \exp\left(-\frac{2n^2 \eta^2}{36\psi^2 \log 1/\eta}\right).$$

Therefore, the above event holds uniformly over all $u \in \mathcal{G}$ with probability at least $1 - 1/(100d^2)$ for our value of $n$. For a non grid point, $u$, consider $v \in \mathcal{G}$ closest to $u$, now we have for all $i$ satisfying $(X_i, Y_i) \in \mathcal{H}$ that:

$$|\langle X_i, u \rangle^2 - \langle X_i, v \rangle^2| \leq 2|\langle X_i, u - v \rangle (X_i, u - v)| + |\langle X_i, u - v \rangle^2| \leq 3\|X_i\|^2 \|u - v\| \leq \eta^4.$$

Therefore, we have for all $\|u\| = 1$:

$$\mathbb{E}_{[n]}[\langle X_i, u \rangle^2 1 \{ (X_i, \epsilon_i) \in \mathcal{H} \land |\langle X_i, u \rangle| \leq 50\psi^2 \log 1/\eta \}] \geq 1 - 2\eta.$$

The lower bound now follows from the fact that removing $10\eta$ points at most reduces the above sum by $500\psi^2 \log 1/\eta$ and the fact that $\mathbb{E}_S[W_i] \geq \mathbb{E}_{[n]}[W_i 1_{\{i \in S\}}]$. \hfill $\square$

In the following lemma, we prove that there exists a good set of data points whose gradients have well behaved covariance structure.

**Lemma C.4.** Let $\mathcal{D}$ satisfy Assumptions 2.4 and 4.4. Then for $n = \tilde{O}(d/\eta^2)$, we have with probability at least $1 - 1/(2d^2)$, that for any $S \subset [n]$ with $|S| \geq (1 - 10\eta)n$:

$$\left\| n^{-1} \sum_{i \in S} G_i(w^*) G_i(w^*)^\top 1 \{ (X_i, \epsilon_i) \in \mathcal{H} \} - \sigma^2 \cdot I \right\| \leq O(\eta \log^2 1/\eta)\sigma^2.$$

**Proof of Lemma C.4.** We first explicitly write out the expression for $G_i(w^*) G_i(w^*)^\top$ as follows:

$$G_i(w^*) G_i(w^*)^\top = \langle X_i, w^* \rangle^2 X_i X_i^\top + 2\epsilon_i \langle X_i, w^* \rangle X_i X_i^\top + \epsilon_i^2 X_i X_i^\top.$$

For the first term in the sum, consider the random matrix $Z = n^{-1} \sum_{i=1}^n Z_i$ where $Z_i = \langle X_i, w^* \rangle^2 X_i X_i^\top 1 \{ (X_i, \epsilon_i) \in \mathcal{H} \}$. We first note that:

$$\mathbb{E}[Z_i^2] = \mathbb{E}[(\langle X_i, w^* \rangle)^4 \|X_i\|^2 X_i X_i^\top 1 \{ (X_i, \epsilon_i) \in \mathcal{H} \}] \leq O(\eta^2 \sigma^4 \log^2 1/\eta) \mathbb{E}[\|X_i\|^4 X_i X_i^\top 1 \{ (X_i, \epsilon_i) \in \mathcal{H} \}] \leq O(\sigma^4 d \eta^2 \log^2 1/\eta) \cdot I,$$
where the first inequality follows from the definition of $\mathcal{H}$ and the last inequality follows from the hypercontractivity of sub-Gaussian distributions and Lemma F.2. Furthermore, we have $\|Z_i\| \leq O(\sigma^2 \eta \log 1 / \eta (d + \log 1 / \eta))$. To bound the expectation of $Z_i$, we have for all $\|u\| = 1$:

$$
\mathbb{E}[u^T Z_i u] \leq \mathbb{E}[\langle X_i, w^* \rangle^2 \langle X_i, u \rangle^2] \leq O(\sigma^2 \eta),
$$

where the last inequality follows from the hypercontractivity of sub-Gaussian distributions and Cauchy-Schwarz. There, we have $\|\mathbb{E}[Z_i]\| \leq O(\sigma^2 \eta)$. Now, applying matrix-Bernstein to $Z$, we get that:

$$
\mathbb{P}\left(\|Z - \mathbb{E}[Z_i]\| \geq \sigma^2 \eta\right) \leq 2d \exp\left(-\frac{n^2 \sigma^4 \eta^2}{C\eta^4 d n^2 \log^2 1/\eta + n \sigma^4 \eta^2}\right) \leq \frac{1}{10d^2},
$$

where the last inequality follows from our setting of $n$. We have now established control over the first term in the expansion of $G_i G_i^\top$.

To control the last term, we will condition on the event from Lemma C.2 which occurs with probability at least $1 - 3 / (10d^2)$.

For the middle term, we first define the random matrix $W_i = \epsilon_i \langle X_i, w^* \rangle X_i X_i^\top 1 \{(X_i, \epsilon_i) \in \mathcal{H}\}$ and consider the random matrix $W = n^{-1} \sum_{i=1}^n W_i$. We first bound the mean of $W_i$. Take any unit vector $u$:

$$
\mathbb{E}[u^T W_i u] = \mathbb{E}[\epsilon_i \langle X_i, w^* \rangle \langle X_i, u \rangle^2 (1 - 1 \{(X_i, \epsilon_i) \not\in \mathcal{H}\})] 
\leq (36 \nu \psi / \sigma^2 \sqrt{\eta} \log 1 / \eta) \mathbb{E}[|\langle X_i, w^* \rangle| 1 \{(X_i, \epsilon_i) \not\in \mathcal{H}\}] 
\leq \eta^4 \sigma^2,
$$

where the last inequality follows from Cauchy-Schwarz, bounds on the moments of sub-Gaussian distributions and $\eta$ being sufficiently small. Proceeding identically to the previous term, we obtain by matrix-Bernstein that:

$$
\mathbb{P}\left(\|W\| \geq \sigma^2 \eta\right) \leq \frac{1}{10d^2}.
$$

Now, for any subset $S \subset [n]$ with $|S| \geq (1 - 10 \eta)n$, we have for any unit vector $u$:

$$
n^{-1} \sum_{i \in S} u^T W_i u = n^{-1} \sum_{i \in S} \epsilon_i \langle X_i, w^* \rangle \langle X_i, u \rangle^2 1 \{(X_i, \epsilon_i) \in \mathcal{H}\}
\leq n^{-1} \sum_{i \in S} \left(\epsilon_i^2 \langle X_i, u \rangle^2 + \langle X_i, w^* \rangle^2 \langle X_i, u \rangle^2\right) 1 \{(X_i, \epsilon_i) \in \mathcal{H}\}
\leq O(\sigma^2 \eta \log 2 / \eta) + O(\sigma^2 \eta) = O(\sigma^2 \eta \log 2 / \eta) + O(\sigma^2 \eta,)
$$

where we used Lemma C.2 for the first term and summed over all elements for the second term. By putting the previous results together, we have:

$$
\left\|n^{-1} \sum_{i \in S} (\epsilon_i^2 X_i X_i^\top + 2 \epsilon_i \langle X_i, w^* \rangle X_i X_i^\top + \langle X_i, w^* \rangle^2 X_i X_i^\top) 1 \{(X_i, \epsilon_i) \in \mathcal{H}\} - \sigma^2 \cdot I\right\|
\leq O(\eta \log 2 / \eta) \sigma^2,
$$

where we used the Lemma C.2 to bound the deviation of the last term from $\sigma^2 \cdot I$ and the results proved in previous displays for the first and second term.

We now present the main result of this section which establishes the deterministic conditions required for the success of Algorithm 3.
Lemma C.5. Let $\mathcal{D}$ satisfy Assumptions 2.4 and 4.4. Then, given $n = \tilde{O}(d/\eta^2)$ 2$\eta$-corrupted samples from $\mathcal{D}$, we have with probability at least $1 - 1/d^2$, there exists a set $S \subset [n]$ with $|S| \geq (1 - 3\eta)n$ such that for any $T \subset S$ with $|T| \geq (1 - 10\eta)n$:

$$
\|\mathbb{E}_T[Y_iX_i] - w^*\| \leq O(1)\sigma\eta \log 1/\eta \text{ and } \|\mathbb{E}_T[Y_i^2X_i^\top] - \sigma^2 \cdot I\| \leq O(1) \cdot \sigma^2 \eta \log^2 1/\eta.
$$

Furthermore, for all $i \in S$, we have $|Y_iX_i| \leq O(1) \cdot \sigma \sqrt{\log 1/\eta} (\sqrt{d} + \sqrt{\log 1/\eta})$.

Proof of Lemma C.5. Assume that the conclusions of Lemmas C.2 to C.4. This happens by the union bound with probability at least $1 - 9/(10d^2)$. Now, consider the set $S' = \{i : (X_i, \epsilon_i) \in \mathcal{H} \}$. Since, we have $\mathbb{P}((X_i, \epsilon_i) \notin \mathcal{H}) \leq \eta^{16}$, we have via the Bernstein bound that with probability at least $1 - d^2/10$ that $|S'| \geq (1 - \eta)n$. Now, since we have $2\eta n$ adversarial corruptions at least $(1 - 3\eta)n$ points of $S'$ are not corrupted. Let $S$ be the uncorrupted points of $S'$. This proves the final claim of the lemma.

Let $T \subset S$ with $|T| \geq (1 - 10\eta)n$. The second claim now follows from the conclusion of Lemma C.4. The first term claim now follows from Lemma C.3 and Assumption 4.4.

This concludes the proof of the lemma.

## D Lower Bound

In this section, we prove that the recovery guarantees for our algorithms are optimal. In both cases, the lower bound follows from similar arguments that for robust mean estimation. Concretely, we prove these lower bounds in the one dimensional case where $X, Y, w^*, \epsilon \in \mathbb{R}$ and in both cases, the distribution over $X$ is such that $X = 1$ with probability 1. It is easy to see that this distribution over $X$ satisfies both Assumptions 2.2 and 2.4. In what follows, we use $d_{TV}$ to denote the total variation distance between two distributions. We first provide the lower bound for the heavy-tailed setting:

**Theorem D.1.** For any $\sigma > 0$, $\eta \in (0, 1)$, there exist two distributions, $\mathcal{D}_1$ and $\mathcal{D}_2$, over pairs $(X, Y)$ with $X, Y \in \mathbb{R}$, satisfying Assumption 2.2 such that:

$$d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{\eta}{2} \text{ and } |w^*_1 - w^*_2| \geq \Omega(\sigma \sqrt{\eta}),$$

where $w^*_1$ and $w^*_2$ are the true parameter vectors for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively.

**Proof.** From Assumption 2.2, we are only required to specify the distributions over $X$ and $\epsilon$. In both cases, we use covariates $X$ such that $\mathbb{P}(X = 1) = 1$. It is clear that $X$ satisfies the L4-L2 hypercontractivity assumption of Assumption 2.2. Now, we design the noise distributions over $\epsilon$ for the two distributions. Using, $\epsilon_1$ and $\epsilon_2$ to denote the noise random variables for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. We define the random variables, $\epsilon_1$ and $\epsilon_2$, as follows:

$$
\mathbb{P}(\epsilon_1 = 0) = 1 \text{ and } \mathbb{P}(\epsilon_2 = x) = \begin{cases}
\frac{\eta}{2}, & \text{if } x = \sigma \cdot \frac{1}{\sqrt{\eta}} \\
1 - \frac{\eta}{2}, & \text{if } x = \sigma \cdot \frac{\sqrt{\eta}}{2(1 - \eta/2)} \\
0, & \text{otherwise.}
\end{cases}
$$

It can be checked that $\epsilon_1$ and $\epsilon_2$ satisfy Assumption 2.2. Finally, we define $w^*_1$ and $w^*_2$. Here, we simply set $w^*_1 = 0$ and $w^*_2 = \sigma \cdot \frac{\sqrt{\eta}}{2(1 - \eta/2)}$. Therefore, the TV distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is the same as the total variation distance between the distributions over the random variables $\epsilon_1$ and $\epsilon_2 + w^*_2$ which is at most $\frac{\eta}{2}$ from the definitions of $\epsilon_1$ and $\epsilon_2$. This concludes the proof of the theorem. \qed
The above proof immediately yields the following corollary which demonstrates that some dependence on the condition number is necessary in the recovery of the parameter vector:

**Corollary D.2.** Let \(d \geq 2\) and \(\kappa > 1\). Then, there exist two distributions, \(\mathcal{D}_1\) and \(\mathcal{D}_2\), over pairs \((X,Y)\) satisfying Assumption 2.2 such that:

\[
d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{\eta}{2} \quad \text{and} \quad |w^*_1 - w^*_2| \geq \Omega(\sigma \sqrt{k\eta})
\]

where \(w^*_1\) and \(w^*_2\) are true parameter vectors for \(\mathcal{D}_1\) and \(\mathcal{D}_2\) respectively.

**Proof.** As in the proof of Theorem D.1, both \(\mathcal{D}_1\) and \(\mathcal{D}_2\) share the same distribution over \(X\). Let \(X = (X_1, \ldots, X_d)\). Now, we will have have

\[
\mathbb{P}(X_i = 1) = 1 \quad \text{for} \quad i \in [d - 1] \quad \text{and} \quad \mathbb{P}(X_d = 1/\sqrt{\kappa}) = 1.
\]

It is easy to see that \(X\) satisfies Assumption 2.2. The distributions over \(\epsilon_1\) and \(\epsilon_2\) are identical to those in Theorem D.1. Finally, to determine \(w^*_1\) and \(w^*_2\), let \(\tilde{w}_1\) and \(\tilde{w}_2\) be the one dimensional parameter values from Theorem D.1. We now set \(w^*_1 = 0\) and \(w^*_2 = (0, \ldots, 0, \tilde{w}_2/\sqrt{\kappa})\). It can again be verified that \(\mathcal{D}_1\) and \(\mathcal{D}_2\) have total variation distance less than \(\eta/2\). From the specific settings of \(w^*_1\) and \(w^*_2\), the proof of the corollary follows. \(\square\)

Through a similar technique, we establish a lower bound for the sub-Gaussian case.

**Theorem D.3.** For any \(\sigma > 0, \eta \in (0,1)\), there exist two distributions, \(\mathcal{D}_1\) and \(\mathcal{D}_2\), over pairs \((X,Y)\) with \(X, Y \in \mathbb{R}\), satisfying Assumption 2.4 such that:

\[
d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leq \frac{\eta}{2} \quad \text{and} \quad |w^*_1 - w^*_2| \geq \Omega(\sigma \eta \sqrt{\log 1/\eta}),
\]

where \(w^*_1\) and \(w^*_2\) are true parameter vectors for \(\mathcal{D}_1\) and \(\mathcal{D}_2\) respectively.

**Proof.** As in the proof of Theorem D.1, we let \(X\) be 1 with probability 1 for both \(\mathcal{D}_1\) and \(\mathcal{D}_2\). Again, \(X\) satisfies the sub-Gaussianity assumptions in Assumption 2.4 with \(\psi = 1\). Deploying the notation from the proof of Theorem D.1, we now design the distribution over \(\epsilon_1\) and \(\epsilon_2\):

\[
\mathbb{P}(\epsilon_1 = 0) = 0 \quad \text{and} \quad \mathbb{P}(\epsilon_2 = x) = \begin{cases} 
\frac{\eta}{2}, & \text{if } x = \sigma \cdot \sqrt{\log 1/\eta} \\
1 - \frac{\eta}{2}, & \text{if } x = \sigma \cdot \frac{\eta \sqrt{\log 1/\eta}}{2(1-\eta/2)} \\
0, & \text{otherwise}.
\end{cases}
\]

Again, we see that both \(\epsilon_1\) and \(\epsilon_2\) satisfy the sub-Gaussianity assumptions in Assumption 2.4. We now finally set \(w^*_1 = 0\) and \(w^*_2 = \sigma \cdot \frac{\eta \sqrt{\log 1/\eta}}{2(1-\eta/2)}\). As before, we have that the TV distance between \(\mathcal{D}_1\) and \(\mathcal{D}_2\) is at most \(\eta/2\) which concludes the proof of the theorem. \(\square\)

**E Fast SDP Solvers**

In this section, we prove the existence of nearly linear time solvers for the class of SDPs required in our algorithms. Our proof follows along the lines of [Cheng et al., 2019] with slight reparametrizations convenient for our analysis. Recall that we aim to solve the following semidefinite program for a given set of points \(Z = \{Z_1, \ldots, Z_n\}\):
\[ \min_{s \in \Delta_{\delta,n}} \lambda_{\max} \left( \sum_{i=1}^{n} s_i Z_i Z_i^\top \right). \]  

(\text{MT})

Cheng et al. [2019] solve this problem via a reduction to the following packing SDP by introducing an additional parameter $\lambda$:

\[
\begin{align*}
&\max_s \sum_{i=1}^{n} s_i \\
\text{Subject to: } &0 \leq s_i \leq \frac{1}{(1 - \delta)n} \\
&\sum_{i=1}^{n} s_i Z_i Z_i^\top \leq \lambda \cdot I.
\end{align*}
\]

(\text{Pack})

Let $\text{OPT}_\delta^*$ denote the optimal value of the program $\text{MT}$, $\text{MT}(\delta)$ denote the program instantiated with $\delta$ and let $\text{Pack}(\delta, \lambda)$ denote the program $\text{Pack}$ instantiated with $\delta, \lambda$ and let $\text{Pack}_*,\lambda$ denote its optimal value. The following quantity is useful throughout the section:

\[ l^* = \min_{s \in \Delta_{\delta,n}} \sum_{i=1}^{n} s_i \|Z_i\|^2. \]

This is simply the average squared lengths of the $(1 - \delta)n$ smallest $Z_i$. We introduce a technical result useful in the following analysis:

**Lemma E.1.** $\text{Pack}(\delta, \text{OPT}_\delta^*)$ has optimal value at least 1.

**Proof.** The lemma follows from the fact that a feasible solution for $\text{MT}(\delta)$ is a feasible solution for $\text{Pack}(\delta, \lambda)$ for $\lambda \geq \text{OPT}_\delta^*$.

The following lemma proves that $l^*$ gives an approximation to $\text{OPT}_\delta^*$ within a factor of $d$.

**Lemma E.2.** The value $l^*$ satisfies:

\[ \text{OPT}_\delta^* \leq l^* \leq d\text{OPT}_\delta^*. \]

**Proof.** The upper bound on $\text{OPT}_\delta^*$ follows from that fact that $\lambda_{\max} Z_i Z_i^\top \leq \|Z_i\|^2$ and the lower bound follows from the inequality $\text{tr} M \leq d\lambda_{\max} M$ for any psd matrix $M$.

In what follows we prove that we can efficiently binary search over the value of $\lambda$ to find a good solution to $\text{MT}$. We refer to $\text{OPT}_\delta$ as the optimal value of $\text{Pack}$ run with $\lambda$. We now show a lemma analogous to Lemma 4.3 in [Cheng et al., 2019] relating to the monotonicity of properties of $\text{OPT}_\lambda$ viewed as a function of $\lambda$.

**Lemma E.3.** The function, $\text{OPT}_\lambda$ when viewed as a function of $\lambda$ is monotonic in $\lambda$.

**Proof.** The lemma follows from the observation that for $\lambda_1 \geq \lambda_2$, a feasible point for $\text{MT}$ with $\lambda_2$ is a feasible point for the program with $\lambda_1$.

We now restate a lemma from [Cheng et al., 2019].

**Lemma E.4.** Given a feasible point for $\text{Pack}(\delta, \lambda)$, $s_i$, with $\sum_{i=1}^{n} s_i \geq 1 - \delta/10$, one can find a feasible solution for $\text{MT}(2\delta)$ with objective value at most $(1 + \delta/2)\lambda$. 

29
The lemma follows by considering the solution \( s'_i = s_i / (\sum_{i=1}^n s_i) \). \( s' \in \Lambda_{2\delta,n} \) as we have \( s'_i \leq \frac{1}{(1-\delta)(1-\delta/10)n} \leq \frac{1}{(1-2\delta)n} \). The bound on the objective value follows from:

\[
\sum_{i=1}^n s'_i Z_i Z_i^\top \leq (1 - \delta/10)^{-1} \sum_{i=1}^n s_i Z_i Z_i^\top \leq (1 + \delta/2)\lambda I.
\]

\( \square \)

We now present a Corollary 4.5 from [Cheng et al., 2019], a consequence of the existence of fast solvers for packing sdps from [Peng et al., 2012]:

**Corollary E.5.** Fix \( 0 < \delta < 1/3 \) and \( \lambda > 0 \). Then, one can obtain in time \( \tilde{O}(nd/\delta^6) \), with probability at least \( 9/10 \), a feasible point for \( \text{Pack}(\delta, \lambda) \), \( \hat{s} \), with:

\[
\sum_{i=1}^n \hat{s}_i \geq (1 - \delta/10)\text{Pack}_{\delta,\lambda}^*.
\]

We conclude with the main lemma of the section.

**Lemma E.6.** Given \( Z = \{Z_i\}_{i=1}^n \) and \( \delta \), one can find with probability at least \( 9/10 \), a feasible solution to \( \text{MT}(2\delta) \) with objective value at most \((1 + \delta)\text{OPT}_{\delta}^*\).

**Proof.** We use a binary search procedure to find an appropriate value of \( \lambda \). To do this, we maintain two indices \( \lambda_l \) and \( \lambda_h \) satisfying the following two conditions:

1. \( \text{OPT}_{\delta}^* \geq \lambda_l \)
2. We have a feasible point \( s_h \) for \( \text{Pack}(\delta, \lambda_h) \) with \( \sum_{i=1}^h s_i \geq (1 - \delta/10) \).

At the beginning we instantiate \( \lambda_l = \lambda^* / d \) and \( \lambda_h = \lambda^* \). The upper bound on \( \lambda^* \) in Lemma E.2 ensures the first condition. For the second one, since \( \lambda^* \geq \text{OPT}_{\delta}^* \), we get from Lemmas E.1 and E.3 and Corollary E.5 the Algorithm from [Peng et al., 2012] obtains a feasible solution for \( \text{Pack}(\delta, \lambda_h) \) with objective value at least \((1 - \delta/10) \). We then run \( \text{Pack} \) with \( \lambda = \lambda_m = (\lambda_l + \lambda_h) / 2 \) and set \( \lambda_h = \lambda_m \) if the objective value is greater than \( 1 - \delta/10 \). Otherwise, we set \( \lambda_l = \lambda_m \). The second condition is trivially maintained while the first condition follows from Lemmas E.1 and E.3 and Corollary E.5. We run the binary search for \( O(\log d/\delta) \) rounds so that we have at the end of the procedure \( \lambda_h - \lambda_l \leq \frac{\delta}{10} \text{OPT}_{\delta}^* \) as we have \( \lambda_h - \lambda_l \leq d\text{OPT}_{\delta}^* \) at the beginning from Lemma E.2. At the end, we return the feasible solution for \( \lambda_h \). The approximation guarantee follows from the fact that \( \lambda_h \leq \lambda_l + \frac{\delta}{10} \text{OPT}_{\delta}^* \leq (1 + \frac{\delta}{10}) \text{OPT}_{\delta}^* \) and Lemma E.4. As in [Cheng et al., 2019], we run the Algorithm \( O(\log \log d/\delta) \) in each step to ensure probability \( 9/10 \) over all rounds. \( \square \)

**F Linear Algebra and Probability Results**

Here we collect the statements (and proofs) of useful results from linear algebra and probability.

**Lemma F.1.** Let \( M \) be a (random) symmetric matrix, then \( \| \mathbb{E}[M^2] \|_2 \leq \| \mathbb{E}[M^2] \|_2 \) and \( \| \mathbb{E}(M - \mathbb{E}[M])^2 \|_2 \leq \| \mathbb{E}[M^2] \|_2 \).

**Proof of Lemma F.1.** Note that \( \mathbb{E}[(M - \mathbb{E}[M])^2] \geq 0 \implies \mathbb{E}[M^2] \geq \mathbb{E}[M]^2 \). Since both matrices are p.s.d. it follows that \( \| \mathbb{E}[M^2] \|_2 \leq \| \mathbb{E}[M^2] \|_2 \). The second claim follows from the variational characterization of the operator norm of the p.s.d. matrix \( (M - \mathbb{E}[M])^2 \). \( \square \)
Lemma F.2. Let $X \sim \mathcal{D}$ be a random vector from a distribution that is $L_4$-L_2 hypercontractive – $\mathbb{E}[\langle v, X \rangle^4] \leq L^2(\mathbb{E}[\langle v, X \rangle^2])^4$ – then

$$||\mathbb{E}[||X||_2^2XX^\top]||_2 \leq L \text{Tr}(\Sigma)||\Sigma||_2.$$ 

Proof of Lemma F.2. We introduce a vector $v$ with $||v||_2 \leq 1$. Then,

$$\mathbb{E}[\langle v, ||X||_2^2XX^\top v \rangle] = \mathbb{E}[||X||_2^2\mathbb{E}[\langle v, X \rangle^2] \leq (\mathbb{E}[||X||_2^2])^{1/2}(\mathbb{E}[\langle v, X \rangle^4])^{1/2},$$

by Cauchy-Schwarz and the Jensen inequality. For the first term we have $(\mathbb{E}[||X||_2^4])^{1/2} \leq \sqrt{L}\text{Tr} \Sigma$ by Lemma F.3. For the second term once again using L4-L2 hypercontractivity we have, $(\mathbb{E}[\langle v, X \rangle^4])^{1/2} \leq \sqrt{L}\mathbb{E}[\langle v, X \rangle^2] \leq \sqrt{L}||\Sigma||_2$. Maximizing over $v$ gives the result. 

□ 

Lemma F.3. Let $X \sim \mathcal{D}$ be a random vector with a distribution that is $L_4$-L_2 hypercontractive. Then,

$$\mathbb{E}[||X||_2^4] \leq L(\text{Tr} \Sigma)^2.$$ 

Proof of Lemma F.3. A short computation using the Cauchy-Schwarz inequality and L4-L2 equivalence shows that,

$$\mathbb{E}[||X||_2^4] = \mathbb{E}[(\sum_{i=1}^d \langle X, e_i \rangle^2)^2] = \mathbb{E}[\sum_{a,b} \langle X, e_a \rangle^2 \langle X, e_b \rangle^2] \leq \sum_{a,b} (\mathbb{E}[\langle X, e_a \rangle^4] \mathbb{E}[\langle X, e_b \rangle^4])^{1/2} \leq L \sum_{a,b} \mathbb{E}[\langle X, e_a \rangle^2] \mathbb{E}[\langle X, e_b \rangle^2] \leq L(\text{Tr} \Sigma)^2.$$ 

□ 

Lemma F.4. Let $X$ be a sub-Gaussian random vector with sub-Gaussian parameter $\phi \geq 1$ and second moment $I$. Then, we have:

$$\mathbb{P}\left(||X|| \geq 7\sqrt{d} + 6\sqrt{\log 1/\delta}\right) \leq \delta.$$ 

Proof of Lemma F.4. Let $\mu$ denote the mean of $X$. We first note that we have $||\mu|| \leq \mathbb{E}[||X||] \leq \sqrt{d}$. Therefore, the mean of $X$ is less than $\sqrt{d}$. Now, pick an 1/2-net over the unit sphere, $\mathcal{G}$. By standard bounds, we have that the number of elements in $\mathcal{G}$ can be upper bounded by $6^d$ Vershynin [2012]. Now, for any $u \in \mathcal{G}$, we have:

$$\mathbb{P}[||u, X - \mu|| \geq t] \leq 2 \exp\left(\frac{-t^2}{2\phi^2}\right).$$

Therefore, $t = 3\sqrt{d + \log 1/\delta}$, and taking a union bound over the at most $O(6^d)$ elements in $\mathcal{G}$, we have with probability $\delta$:

$$||X|| \leq ||\mu|| + 2 \max_{u \in \mathcal{G}} \langle X, u \rangle \leq 7\sqrt{d} + 6\sqrt{\log 1/\delta}.$$ 

□ 

Lemma F.5. Let $X$ be a sub-Gaussian random variable with second moment 1 and sub-Gaussianity parameter $\phi$. Then, we have for sufficiently small $\eta$:

$$\mathbb{E}[X^41\{|X| > 6\phi \sqrt{\log 1/\eta}\}] \leq \phi^4 \eta^8.$$
Proof of Lemma F.5. Since we have $|\mathbb{E}[X]| \leq \mathbb{E}[|X|] \leq 1$. Now, we get:

$$
\mathbb{E}[X^4 1 \left\{ |X| > 6\phi \sqrt{\log 1/\eta} \right\}] \leq (\mathbb{E}[X^8])^{1/2} (\mathbb{P}[|X| > 6\phi \sqrt{\log 1/\eta}])^{1/2} \leq \phi^4 \eta^8.
$$

\[\square\]

Lemma F.6. Assume the setting of Lemma F.5. Then, we have for all $k$:

$$
\mathbb{E}[X^k] \leq k^{1/2} (C\phi)^k,
$$

for some absolute constant $C$.

Proof of Lemma F.6. The result follows from applying the result of Lemma 5.5 to the centered sub-Gaussian random variable $Y = X - \mathbb{E}[X]$ and from the fact that $\mathbb{E}[(Y + \mathbb{E}[X])^k] \leq 2^k \mathbb{E}[Y^k + \mathbb{E}[X]^k]$ and noting that $\mathbb{E}[|X|] \leq 1$ as in the proof of Lemma F.5.

Lemma F.7. Let $n$ be an integer and let $\Delta_{\delta,n}$ be the set of distributions over $[n]$ with $0 \leq s_i \leq \frac{1}{(1 - \delta)n}$ for all $s \in \Delta_\delta$. Then, the points $E_\delta = \{\{\alpha_i\}_{i=1}^n : \alpha_i = 1/((1 - \delta)n) \text{ for } (1 - \delta)n \text{ elements}\}$ form the extreme points of $\Delta_\delta$.

Proof of Lemma F.7. Note that $\Delta_\delta$ is a polytope in $n$ dimension and all points in $E_\delta$ satisfy at least $d$ linearly independent constraints of the polytope. Therefore, all the points in $E_\delta$ are extreme points. To show that $E_\delta$ contains all extreme points, consider an extreme point $\alpha$ not in $E_\delta$. Then, there exists $i$ such that $\alpha_i \neq 0, 1/((1 - \delta)n)$. Since $\delta n$ is an integer, there exists $j \neq i$ such that $\alpha_j \neq 0, 1/((1 - \delta)n)$. Therefore, for small $\xi$, we have $\alpha'$ defined by $\alpha'_k = \alpha_k$ for all $k \neq i$, $j$ and $\alpha'_i = \alpha_i + \xi$ and $\alpha'_j = \alpha_j - \xi$ is in $\Delta_{\delta,n}$ which contradicts $\alpha$ being an extreme point. Therefore, $E_\delta$ contains all extreme points.

Lemma F.8. Assume the setting of Lemma F.7 for $\delta_1$ and $\delta_2$. Then, we have for any $\alpha \in \Delta_{\delta_1,n}$ and $\beta \in \Delta_{\delta_2,n}$:

$$
\text{Dist}_{TV}(\alpha, \beta) \leq \delta_1 + \delta_2.
$$

Proof of Lemma F.8. We prove that $\text{Dist}_{TV}(\alpha, \text{Unif}([n])) \leq \delta_1$. This follows from the fact that maximizing the total variation distance from Unif([n]) is a linear program over $\Delta_{\delta,n}$ and it is maximized at the extreme points described by Lemma F.7. The result follows from a similar proof for $\beta$ and triangle inequality. \[\square\]