Robust Regression via Hard Thresholding

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of Robust Least Squares Regression (RLSR) where several response variables can be adversarially corrupted. More specifically, for a data matrix $X \in \mathbb{R}^{p \times n}$ and an underlying model $\mathbf{w}^*$, the response vector is generated as $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}$ where $\mathbf{b} \in \mathbb{R}^n$ is the corruption vector supported over at most $C \cdot n$ coordinates. Existing exact recovery results for RLSR focus solely on L_1-penalty based convex formulations and impose relatively strict model assumptions such as requiring the corruptions $\mathbf{b}$ to be selected independently of X.

In this work, we study a simple hard-thresholding algorithm called TORRENT which, under mild conditions on X, can recover $\mathbf{w}^*$ exactly even if $\mathbf{b}$ corrupts the response variables in an *adversarial* manner, i.e. both the support and entries of $\mathbf{b}$ are selected adversarially after observing X and $\mathbf{w}^*$. Our results hold under *deterministic* assumptions which are satisfied if X is sampled from any sub-Gaussian distribution. Finally unlike existing results that apply only to a fixed $\mathbf{w}^*$, generated independently of X, our results are *universal* and hold for any $\mathbf{w}^* \in \mathbb{R}^p$.

Next, we propose gradient descent-based extensions of TORRENT that can scale efficiently to large scale problems, such as high dimensional sparse recovery. and prove similar recovery guarantees for these extensions. Empirically we find TORRENT, and more so its extensions, offering significantly faster recovery than the state-of-the-art L_1 solvers. For instance, even on moderate-sized datasets (with $p = 50K$) with around 40% corrupted responses, a variant of our proposed method called TORRENT-HYB is more than $20\times$ faster than the best L_1 solver.

"If among these errors are some which appear too large to be admissible, then those equations which produced these errors will be rejected, as coming from too faulty experiments, and the unknowns will be determined by means of the other equations, which will then give much smaller errors."
A. M. Legendre, *On the Method of Least Squares.* 1805.

1 Introduction

Robust Least Squares Regression (RLSR) addresses the problem of learning a reliable set of regression coefficients in the presence of several arbitrary corruptions in the *response* vector. Owing to the wide-applicability of regression, RLSR features as a critical component of several important real-world applications in a variety of domains such as signal processing [1], economics [2], computer vision [3, 4], and astronomy [2].

Given a data matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with n data points in $\mathbb{R}^p$ and the corresponding response vector $\mathbf{y} \in \mathbb{R}^n$, the goal of RLSR is to learn a $\hat{\mathbf{w}}$ such that,

$$(\hat{\mathbf{w}}, \hat{S}) = \underset{\substack{\mathbf{w} \in \mathbb{R}^p \\ S \subset [n] : |S| \geq (1-\beta)\cdot n}}{\arg\min} \sum_{i \in S} (y_i - \mathbf{x}_i^T \mathbf{w})^2, \tag{1}$$

That is, we wish to simultaneously determine the set of corruption free points $\hat{S}$ and also estimate the best model parameters over the set of clean points. However, the optimization problem given above is non-convex (jointly in $\mathbf{w}$ and S) in general and might not directly admit efficient solutions. Indeed there exist reformulations of this problem that are known to be NP-hard to optimize [1].

To address this problem, most existing methods with provable guarantees assume that the observations are obtained from some generative model. A commonly adopted model is the following

$$\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}, \tag{2}$$

where $\mathbf{w}^* \in \mathbb{R}^p$ is the *true* model vector that we wish to estimate and $\mathbf{b} \in \mathbb{R}^n$ is the corruption vector that can have arbitrary values. A common assumption is that the corruption vector is *sparsely* supported i.e. $\|\mathbf{b}\|_0 \le \alpha \cdot n$ for some $\alpha > 0$.

Recently, [4] and [5] obtained a surprising result which shows that one can recover $\mathbf{w}^*$ *exactly* even when $\alpha \lesssim 1$, i.e., when almost all the points are corrupted, by solving an L_1-penalty based convex optimization problem: $\min_{\mathbf{w},\mathbf{b}} \|\mathbf{w}\|_1 + \lambda \|\mathbf{b}\|_1$, s.t., $X^\top\mathbf{w} + \mathbf{b} = \mathbf{y}$. However, these results require the corruption vector $\mathbf{b}$ to be selected oblivious of X and $\mathbf{w}^*$. Moreover, the results impose severe restrictions on the data distribution, requiring that the data be either sampled from an isotropic Gaussian ensemble [4], or row-sampled from an incoherent orthogonal matrix [5]. Finally, these results hold only for a fixed $\mathbf{w}^*$ and are not universal in general.

In contrast, [6] studied RLSR with less stringent assumptions, allowing arbitrary corruptions in response variables as well as in the data matrix X, and proposed a trimmed inner product based algorithm for the problem. However, their recovery guarantees are significantly weaker. Firstly, they are able to recover $\mathbf{w}^*$ only upto an additive error $\alpha\sqrt{p}$ (or $\alpha\sqrt{s}$ if $\mathbf{w}^*$ is s-sparse). Hence, they require $\alpha \le 1/\sqrt{p}$ just to claim a non-trivial bound. Note that this amounts to being able to tolerate only a vanishing fraction of corruptions. More importantly, even with $n \to \infty$ and extremely small α they are unable to guarantee exact recovery of $\mathbf{w}^*$. A similar result was obtained by [7], albeit using a sub-sampling based algorithm with stronger assumptions on $\mathbf{b}$.

In this paper, we focus on a simple and natural thresholding based algorithm for RLSR. At a high level, at each step t, our algorithm alternately estimates an *active set* S_t of "clean" points and then updates the model to obtain $\mathbf{w}^{t+1}$ by minimizing the least squares error on the active set. This intuitive algorithm seems to embody a long standing heuristic first proposed by Legendre [8] over two centuries ago (see introductory quotation in this paper) that has been adopted in later literature [9, 10] as well. However, to the best of our knowledge, this technique has never been rigorously analyzed before in non-asymptotic settings, despite its appealing simplicity.

Our Contributions: The main contribution of this paper is an exact recovery guarantee for the thresholding algorithm mentioned above that we refer to as TORRENT-FC (see Algorithm 1). We provide our guarantees in the model given in 2 where the corruptions $\mathbf{b}$ are selected *adversarially* but restricted to have at most $\alpha \cdot n$ non-zero entries where $\alpha < 1/2$ is a global constant dependent only on X[1]. Under *deterministic* conditions on X, namely the subset strong convexity (SSC) and smoothness (SSS) properties (see Definition 1), we guarantee that TORRENT-FC converges at a *geometric* rate and recovers $\mathbf{w}^*$ exactly. We further show that these properties (SSC and SSS) are satisfied w.h.p. if a) the data X is sampled from a sub-Gaussian distribution and, b) $n \ge p\log p$.

We would like to stress three key advantages of our result over the results of [4, 5]: a) we allow $\mathbf{b}$ to be adversarial, i.e., both support and values of $\mathbf{b}$ to be selected adversarially based on X and $\mathbf{w}^*$, b) we make assumptions on data that are natural, as well as significantly less restrictive than what existing methods make, and c) our analysis admits universal guarantees, i.e., holds for *any* $\mathbf{w}^*$.

We would also like to stress that while hard-thresholding based methods have been studied rigorously for the sparse-recovery problem [11, 12], hard-thresholding has not been studied formally for the robust regression problem. Moreover, the two problems are completely different and hence techniques from sparse-recovery analysis do not extend to robust regression.

Despite its simplicity, TORRENT-FC does not scale very well to datasets with large p as it solves least squares problems at each iteration. We address this issue by designing a gradient descent

[1]Note that for an adaptive adversary, as is the case in our work, recovery cannot be guaranteed for $\alpha \ge 1/2$ since the adversary can introduce corruptions as $\mathbf{b}_i = \mathbf{x}_i^\top(\widetilde{\mathbf{w}} - \mathbf{w}^*)$ for an adversarially chosen model $\widetilde{\mathbf{w}}$. This would make it impossible for any algorithm to distinguish between $\mathbf{w}^*$ and $\widetilde{\mathbf{w}}$ thus making recovery impossible.

based algorithm (TORRENT-GD), and a hybrid algorithm (TORRENT-Hyb), both of which enjoy a geometric rate of convergence and can recover $\mathbf{w}^*$ under the model assumptions mentioned above. We also propose extensions of TORRENT for the RLSR problem in the sparse regression setting where $p \gg n$ but $\|\mathbf{w}^*\|_0 = s^* \ll p$. Our algorithm TORRENT-HD is based on TORRENT-FC but uses the Iterative Hard Thresholding (IHT) algorithm, a popular algorithm for sparse regression. As before, we show that TORRENT-HD also converges geometrically to $\mathbf{w}^*$ if a) the corruption index α is less than some constant C, b) X is sampled from a sub-Gaussian distribution and, c) $n \geq s^* \log p$.

Finally, we experimentally evaluate existing L_1-based algorithms and our hard thresholding-based algorithms. The results demonstrate that our proposed algorithms (TORRENT-(FC/GD/HYB)) can be significantly faster than the best L_1 solvers, exhibit better recovery properties, as well as be more robust to dense white noise. For instance, on a problem with $50K$ dimensions and 40% corruption, TORRENT-HYB was found to be $20\times$ faster than L_1 solvers, as well as achieve lower error rates.

Paper Organization: We give a formal definition of the RLSR problem in the next section. We then introduce our family of algorithms in Section 3 and prove their convergence guarantees in Section 4. We present extensions to sparse robust regression in Section 5 and empirical results in Section 6.

2 Problem Formulation

Given a set of data points $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbb{R}^p$ and the corresponding response vector $\mathbf{y} \in \mathbb{R}^n$, the goal is to recover a parameter vector $\mathbf{w}^*$ which solves the RLSR problem (1). We assume that the response vector $\mathbf{y}$ is generated using the following model:

$$\mathbf{y} = \mathbf{y}^* + \mathbf{b} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{y}^* = X^\top \mathbf{w}^*.$$

Hence, in the above model, (1) reduces to estimating $\mathbf{w}^*$. We allow the model $\mathbf{w}^*$ representing the regressor, to be chosen in an adaptive manner *after* the data features have been generated.

The above model allows two kinds of perturbations to y_i – dense but bounded noise ε_i (e.g. white noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \sigma \geq 0$), as well as potentially unbounded corruptions b_i – to be introduced by an adversary. The only requirement we enforce is that the gross corruptions be sparse. $\boldsymbol{\varepsilon}$ shall represent the dense noise vector, for example $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_{n\times n})$, and $\mathbf{b}$, the corruption vector such that $\|\mathbf{b}\|_0 \leq \alpha \cdot n$ for some *corruption index* $\alpha > 0$. We shall use the notation $S_* = \overline{\text{supp}(\mathbf{b})} \subseteq [n]$ to denote the set of "clean" points, i.e. points that have not faced unbounded corruptions. We consider adaptive adversaries that are able to view the generated data points $\mathbf{x}_i$, as well as the clean responses y_i^* and dense noise values ε_i before deciding which locations to corrupt and by what amount.

We denote the unit sphere in p dimensions using S^{p-1}. For any $\gamma \in (0, 1]$, we let $\mathcal{S}_\gamma = \{S \subset [n] : |S| = \gamma \cdot n\}$ denote the set of all subsets of size $\gamma \cdot n$. For any set S, we let $X_S := [\mathbf{x}_i]_{i\in S} \in \mathbb{R}^{p\times|S|}$ denote the matrix whose columns are composed of points in that set. Also, for any vector $\mathbf{v} \in \mathbb{R}^n$ we use the notation $\mathbf{v}_S$ to denote the $|S|$-dimensional vector consisting of those components that are in S. We use $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ to denote, respectively, the smallest and largest eigenvalues of a square symmetric matrix X. We now introduce two properties, namely, *Subset Strong Convexity* and *Subset Strong Smoothness*, which are key to our analyses.

Definition 1 (SSC and SSS Properties)**.** *A matrix $X \in \mathbb{R}^{p\times n}$ satisfies the* Subset Strong Convexity Property *(resp.* Subset Strong Smoothness Property*) at level γ with strong convexity constant λ_γ (resp. strong smoothness constant Λ_γ) if the following holds:*

$$\lambda_\gamma \leq \min_{S\in\mathcal{S}_\gamma} \lambda_{\min}(X_S X_S^\top) \leq \max_{S\in\mathcal{S}_\gamma} \lambda_{\max}(X_S X_S^\top) \leq \Lambda_\gamma.$$

Remark 1. We note that the uniformity enforced in the definitions of the SSC and SSS properties is not for the sake of convenience but rather a necessity. Indeed, a uniform bound is required in face of an adversary which can perform corruptions *after* data and response variables have been generated, and choose to corrupt precisely that set of points where the SSC and SSS parameters are the worst.

3 TORRENT: Thresholding Operator-based Robust Regression Method

We now present TORRENT, a Thresholding Operator-based Robust RegrEssioN meTHod for performing robust regression at scale. Key to our algorithms is the *Hard Thresholding Operator* which we define below.

Algorithm 1 TORRENT: Thresholding Operator-based Robust RegrEssioN meThod

Input: Training data $\{\mathbf{x}_i, y_i\}, i = 1 \ldots n$, step length η, thresholding parameter β, tolerance ϵ

$\mathbf{w}^0 \leftarrow \mathbf{0}, S_0 = [n], t \leftarrow 0, \mathbf{r}^0 \leftarrow \mathbf{y}$
while $\left\|\mathbf{r}^t_{S_t}\right\|_2 > \epsilon$ **do**
  $\mathbf{w}^{t+1} \leftarrow \text{UPDATE}(\mathbf{w}^t, S_t, \eta, \mathbf{r}^t, S_{t-1})$
  $r_i^{t+1} \leftarrow \left(y_i - \left\langle \mathbf{w}^{t+1}, \mathbf{x}_i \right\rangle\right)$
  $S_{t+1} \leftarrow \text{HT}(\mathbf{r}^{t+1}, (1-\beta)n)$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}^t$

Algorithm 2 TORRENT-FC

Input: Current model $\mathbf{w}$, current active set S

return $\arg\min_{\mathbf{w}} \sum_{i \in S} \left(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle\right)^2$

Algorithm 3 TORRENT-GD

Input: Current model $\mathbf{w}$, current active set S, step size η

$\mathbf{g} \leftarrow X_S(X_S^\top \mathbf{w} - \mathbf{y}_S)$
return $\mathbf{w} - \eta \cdot \mathbf{g}$

Algorithm 4 TORRENT-HYB

Input: Current model $\mathbf{w}$, current active set S, step size η, current residuals $\mathbf{r}$, previous active set S'

// `Use the GD update if the active set` S `is changing a lot`
if $|S \backslash S'| > \Delta$ **then**
  $\mathbf{w}' \leftarrow \text{UPDATE-GD}(\mathbf{w}, S, \eta, \mathbf{r}, S')$
else
// `If stable, use the FC update`
  $\mathbf{w}' \leftarrow \text{UPDATE-FC}(\mathbf{w}, S)$
end if
return $\mathbf{w}'$

Definition 2 (Hard Thresholding Operator). *For any vector $\mathbf{v} \in \mathbb{R}^n$, let $\sigma_{\mathbf{v}} \in S_n$ be the permutation that orders elements of $\mathbf{v}$ in ascending order of their magnitudes i.e. $\left|\mathbf{v}_{\sigma_{\mathbf{v}}(1)}\right| \leq \left|\mathbf{v}_{\sigma_{\mathbf{v}}(2)}\right| \leq \ldots \leq \left|\mathbf{v}_{\sigma_{\mathbf{v}}(n)}\right|$. Then for any $k \leq n$, we define the hard thresholding operator as*

$$\text{HT}(\mathbf{v}; k) = \left\{ i \in [n] : \sigma_{\mathbf{v}}^{-1}(i) \leq k \right\}$$

Using this operator, we present our algorithm TORRENT (Algorithm 1) for robust regression. TORRENT follows a most natural iterative strategy of, alternately, estimating an *active set* of points which have the least residual error on the current regressor, and then updating the regressor to provide a better fit on this active set. We offer three variants of our algorithm, based on how aggressively the algorithm tries to fit the regressor to the current active set.

We first propose a fully corrective algorithm TORRENT-FC (Algorithm 2) that performs a fully corrective least squares regression step in an effort to minimize the regression error on the active set. This algorithm makes significant progress in each step, but at a cost of more expensive updates. To address this, we then propose a milder, gradient descent-based variant TORRENT-GD (Algorithm 3) that performs a much cheaper update of taking a single step in the direction of the gradient of the objective function on the active set. This reduces the regression error on the active set but does not minimize it. This turns out to be beneficial in situations where dense noise is present along with sparse corruptions since it prevents the algorithm from overfitting to the current active set.

Both the algorithms proposed above have their pros and cons – the FC algorithm provides significant improvements with each step, but is expensive to execute whereas the GD variant, although efficient in executing each step, offers slower progress. To get the best of both these algorithms, we propose a third, hybrid variant TORRENT-HYB (Algorithm 4) that adaptively selects either the FC or the GD update depending on whether the active set is stable across iterations or not.

In the next section we show that this hard thresholding-based strategy offers a linear convergence rate for the algorithm in all its three variations. We shall also demonstrate the applicability of this technique to high dimensional sparse recovery settings in a subsequent section.

4 Convergence Guarantees

For the sake of ease of exposition, we will first present our convergence analyses for cases where dense noise is not present i.e. $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$ and will handle cases with dense noise *and* sparse corruptions later. We first analyze the fully corrective TORRENT-FC algorithm. The convergence proof in this case relies on the optimality of the two steps carried out by the algorithm, the fully corrective step that selects the best regressor on the active set, and the hard thresholding step that discovers a new active set by selecting points with the least residual error on the current regressor.

Theorem 3. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix and $\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b}$ be the corrupted output with $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Let Algorithm 2 be executed on this data with the thresholding parameter set to $\beta \geq \alpha$. Let Σ_0 be an invertible matrix such that $\widetilde{X} = \Sigma_0^{-1/2} X$ satisfies the SSC and SSS properties at level γ with constants λ_γ and Λ_γ respectively (see Definition 1). If the data satisfies $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < 1$, then after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 2 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq \epsilon$.*

Proof (Sketch). Let $\mathbf{r}^t = \mathbf{y} - X^\top \mathbf{w}^t$ be the vector of residuals at time t and $C_t = X_{S_t} X_{S_t}^\top$. Also let $S_* = \overline{\text{supp}(\mathbf{b})}$ be the set of uncorrupted points. The fully corrective step ensures that

$$\mathbf{w}^{t+1} = C_t^{-1} X_{S_t} \mathbf{y}_{S_t} = C_t^{-1} X_{S_t} \left(X_{S_t}^\top \mathbf{w}^* + \mathbf{b}_{S_t}\right) = \mathbf{w}^* + C_t^{-1} X_{S_t} \mathbf{b}_{S_t},$$

whereas the hard thresholding step ensures that $\left\|\mathbf{r}^{t+1}_{S_{t+1}}\right\|_2^2 \leq \left\|\mathbf{r}^{t+1}_{S_*}\right\|_2^2$. Combining the two gives us

$$\begin{aligned}
\left\|\mathbf{b}_{S_{t+1}}\right\|_2^2 &\leq \left\|X_{S_* \backslash S_{t+1}}^\top C_t^{-1} X_{S_t} \mathbf{b}_{S_t}\right\|_2^2 + 2 \cdot \mathbf{b}_{S_{t+1}}^\top X_{S_{t+1}}^\top C_t^{-1} X_{S_t} \mathbf{b}_{S_t} \\
&\overset{\zeta_1}{=} \left\|\widetilde{X}_{S_* \backslash S_{t+1}}^\top \left(\widetilde{X}_{S_t} \widetilde{X}_{S_t}^T\right)^{-1} \widetilde{X}_{S_t} \mathbf{b}_{S_t}\right\|_2^2 + 2 \cdot \mathbf{b}_{S_{t+1}}^\top \widetilde{X}_{S_{t+1}}^\top \left(\widetilde{X}_{S_t} \widetilde{X}_{S_t}^T\right)^{-1} \widetilde{X}_{S_t} \mathbf{b}_{S_t} \\
&\overset{\zeta_2}{\leq} \frac{\Lambda_\beta^2}{\lambda_{1-\beta}^2} \cdot \left\|\mathbf{b}_{S_t}\right\|_2^2 + 2 \cdot \frac{\Lambda_\beta}{\lambda_{1-\beta}} \cdot \left\|\mathbf{b}_{S_t}\right\|_2 \left\|\mathbf{b}_{S_{t+1}}\right\|_2,
\end{aligned}$$

where ζ_1 follows from setting $\widetilde{X} = \Sigma_0^{-1/2} X$ and $X_S^\top C_t^{-1} X_{S'} = \widetilde{X}_S^\top (\widetilde{X}_{S_t} \widetilde{X}_{S_t}^\top)^{-1} \widetilde{X}_{S'}$ and ζ_2 follows from the SSC and SSS properties, $\left\|\mathbf{b}_{S_t}\right\|_0 \leq \|\mathbf{b}\|_0 \leq \beta \cdot n$ and $|S_* \backslash S_{t+1}| \leq \beta \cdot n$. Solving the quadratic equation and performing other manipulations gives us the claimed result. □

Theorem 3 relies on a deterministic (*fixed design*) assumption, specifically $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < 1$ in order to guarantee convergence. We can show that a large class of random designs, including Gaussian and sub-Gaussian designs actually satisfy this requirement. That is to say, data generated from these distributions satisfy the SSC and SSS conditions such that $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < 1$ with high probability. Theorem 4 explicates this for the class of Gaussian designs.

Theorem 4. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix with each $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Let $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$ and $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Also, let $\alpha \leq \beta < \frac{1}{65}$ and $n \geq \Omega\left(p + \log \frac{1}{\delta}\right)$. Then, with probability at least $1-\delta$, the data satisfies $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < \frac{9}{10}$. More specifically, after $T \geq 10 \log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)$ iterations of Algorithm 1 with the thresholding parameter set to β, we have $\left\|\mathbf{w}^T - \mathbf{w}^*\right\| \leq \epsilon$.*

Remark 2. Note that Theorem 4 provides rates that are independent of the condition number $\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ of the distribution. We also note that results similar to Theorem 4 can be proven for the larger class of sub-Gaussian distributions. We refer the reader to Section G for the same.

Remark 3. We remind the reader that our analyses can readily accommodate dense noise in addition to sparse unbounded corruptions. We direct the reader to Appendix A which presents convergence proofs for our algorithms in these settings.

Remark 4. We would like to point out that the design requirements made by our analyses are very mild when compared to existing literature. Indeed, the work of [4] assumes the *Bouquet Model* where distributions are restricted to be isotropic Gaussians whereas the work of [5] assumes a more stringent model of sub-orthonormal matrices, something that even Gaussian designs do not satisfy. Our analyses, on the other hand, hold for the general class of sub-Gaussian distributions.

We now analyze the TORRENT-GD algorithm which performs cheaper, gradient-style updates on the active set. We will show that this method nevertheless enjoys a linear rate of convergence.

Theorem 5. *Let the data settings be as stated in Theorem 3 and let Algorithm 3 be executed on this data with the thresholding parameter set to $\beta \geq \alpha$ and the step length set to $\eta = \frac{1}{\Lambda_{1-\beta}}$. If the data* *satisfies* $\max\left\{\eta\sqrt{\Lambda_\beta}, 1-\eta\lambda_{1-\beta}\right\} \leq \frac{1}{4}$, *then after* $t = \mathcal{O}\left(\log\left(\frac{\|b\|_2}{\sqrt{n}}\frac{1}{\epsilon}\right)\right)$ *iterations, Algorithm 1 obtains an* ϵ*-accurate solution* $\mathbf{w}^t$ *i.e.* $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$.

Similar to TORRENT-FC, the assumptions made by the TORRENT-GD algorithm are also satisfied by the class of sub-Gaussian distributions. The proof of Theorem 5, given in Appendix D, details these arguments. Given the convergence analyses for TORRENT-FC and GD, we now move on to provide a convergence analysis for the hybrid TORRENT-HYB algorithm which interleaves FC and GD steps. Since the exact interleaving adopted by the algorithm depends on the data, and not known in advance, this poses a problem. We address this problem by giving below a uniform convergence guarantee, one that applies to *every interleaving* of the FC and GD update steps.

Theorem 6. *Suppose Algorithm 4 is executed on data that allows Algorithms 2 and 3 a convergence rate of* η_{FC} *and* η_{GD} *respectively. Suppose we have* $2\cdot\eta_{\text{FC}}\cdot\eta_{\text{GD}} < 1$. *Then for* any *interleavings of the FC and GD steps that the policy may enforce, after* $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ *iterations, Algorithm 4 ensures an* ϵ*-optimal solution i.e.* $\|\mathbf{w}^t - \mathbf{w}^*\| \leq \epsilon$.

We point out to the reader that the assumption made by Theorem 6 i.e. $2\cdot\eta_{\text{FC}}\cdot\eta_{\text{GD}} < 1$ is readily satisfied by random sub-Gaussian designs, albeit at the cost of reducing the noise tolerance limit. As we shall see, TORRENT-HYB offers attractive convergence properties, merging the fast convergence rates of the FC step, as well as the speed and protection against overfitting provided by the GD step.

5 High-dimensional Robust Regression

In this section, we extend our approach to the robust high-dimensional sparse recovery setting. As before, we assume that the response vector $\mathbf{y}$ is obtained as: $\mathbf{y} = X^\top\mathbf{w}^* + \mathbf{b}$, where $\|\mathbf{b}\|_0 \leq \alpha\cdot n$. However, this time, we also assume that $\mathbf{w}^*$ is s^*-sparse i.e. $\|\mathbf{w}^*\|_0 \leq s^*$. As before, we shall neglect white/dense noise for the sake of simplicity. We reiterate that it is not possible to use existing results from sparse recovery (such as [11, 12]) directly to solve this problem.

Our objective would be to recover a *sparse* model $\hat{\mathbf{w}}$ so that $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq \epsilon$. The challenge here is to forgo a sample complexity of $n \gtrsim p$ and instead, perform recovery with $n \sim s^*\log p$ samples alone. For this setting, we modify the FC update step of TORRENT-FC method to the following:

$$\mathbf{w}^{t+1} \leftarrow \inf_{\|\mathbf{w}\|_0 \leq s} \sum_{i\in S_t} \left(y_i - \langle\mathbf{w}, \mathbf{x}_i\rangle\right)^2, \tag{3}$$

for some *target* sparsity level $s \ll p$. We refer to this modified algorithm as TORRENT-HD. Assuming X satisfies the RSC/RSS properties (defined below), (3) can be solved efficiently using results from sparse recovery (for example the IHT algorithm [11, 13] analyzed in [12]).

Definition 7 (RSC and RSS Properties)**.** *A matrix* $X \in \mathbb{R}^{p\times n}$ *will be said to satisfy the* Restricted Strong Convexity Property *(resp.* Restricted Strong Smoothness Property*) at level* $s = s_1 + s_2$ *with strong convexity constant* $\alpha_{s_1+s_2}$ *(resp. strong smoothness constant* $L_{s_1+s_2}$*) if the following holds for all* $\|\mathbf{w}_1\|_0 \leq s_1$ *and* $\|\mathbf{w}_2\|_0 \leq s_2$*:*

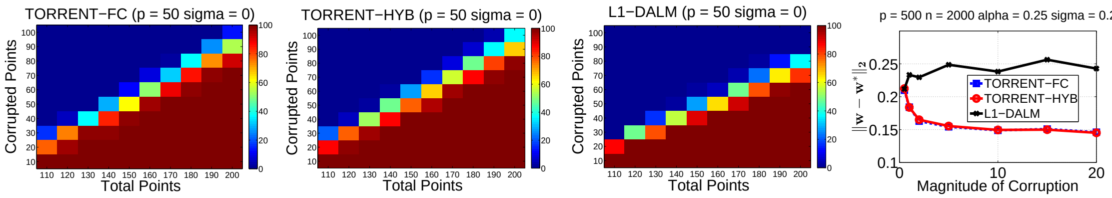
$$\alpha_s\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 \leq \left\|X^\top(\mathbf{w}_1 - \mathbf{w}_2)\right\|_2^2 \leq L_s\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$$

For our results, we shall require the subset versions of both these properties.

Definition 8 (SRSC and SRSS Properties)**.** *A matrix* $X \in \mathbb{R}^{p\times n}$ *will be said to satisfy the* Subset Restricted Strong Convexity *(resp.* Subset Restricted Strong Smoothness*) Property at level* (γ, s) *with strong convexity constant* $\alpha_{(\gamma,s)}$ *(resp. strong smoothness constant* $L_{(\gamma,s)}$*) if for all subsets* $S \in \mathcal{S}_\gamma$*, the matrix* X_S *satisfies the RSC (resp. RSS) property at level* s *with constant* α_s *(resp.* L_s*).*

We now state the convergence result for the TORRENT-HD algorithm.

Theorem 9. *Let* $X \in \mathbb{R}^{p\times n}$ *be the given data matrix and* $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}$ *be the corrupted output with* $\|\mathbf{w}^*\|_0 \leq s^*$ *and* $\|\mathbf{b}\|_0 \leq \alpha\cdot n$. *Let* Σ_0 *be an invertible matrix such that* $\Sigma_0^{-1/2}X$ *satisfies the SRSC and SRSS properties at level* $(\gamma, 2s+s^*)$ *with constants* $\alpha_{(\gamma,2s+s^*)}$ *and* $L_{(\gamma,2s+s^*)}$ *respectively (see Definition 8). Let Algorithm 2 be executed on this data with the* TORRENT-HD *update, thresholding parameter set to* $\beta \geq \alpha$*, and* $s \geq 32\left(\frac{L_{(1-\beta,2s+s^*)}}{\alpha_{(1-\beta,2s+s^*)}}\right)$.



(a) (b) (c) (d)

Figure 1: (a), (b) and (c) Phase-transition diagrams depicting the recovery properties of the TORRENT-FC, TORRENT-HYB and L_1 algorithms. The colors red and blue represent a high and low probability of success resp. A method is considered successful in an experiment if it recovers $\mathbf{w}^*$ upto a 10^{-4} relative error. Both variants of TORRENT can be seen to recover $\mathbf{w}^*$ in presence of larger number of corruptions than the L_1 solver. (d) Variation in recovery error with the magnitude of corruption. As the corruption is increased, TORRENT-FC and TORRENT-HYB show improved performance while the problem becomes more difficult for the L_1 solver.

If X also satisfies $\frac{4L_{(\beta,s+s^*)}}{\alpha_{(1-\beta,s+s^*)}} < 1$, then after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 2 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$.

In particular, if X is sampled from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ and $n \geq \Omega\left(s^* \cdot \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \log p\right)$, then for all values of $\alpha \leq \beta < \frac{1}{65}$, we can guarantee $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$ after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations of the algorithm (w.p. $\geq 1 - 1/n^{10}$).

Remark 5. The sample complexity required by Theorem 9 is identical to the one required by analyses for high dimensional sparse recovery [12], save constants. Also note that TORRENT-HD can tolerate the same corruption index as TORRENT-FC.

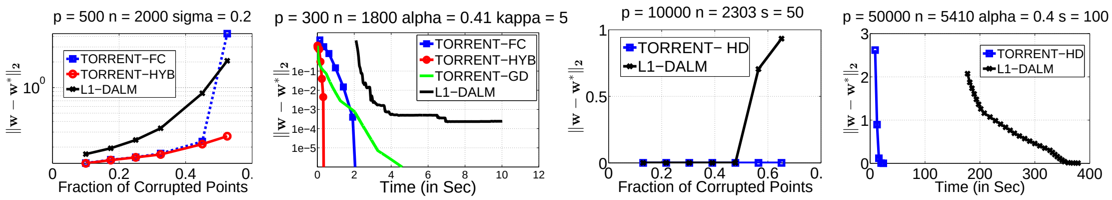
6 Experiments

Several numerical simulations were carried out on linear regression problems in low-dimensional, as well as sparse high-dimensional settings. The experiments show that TORRENT not only offers statistically better recovery properties as compared to L_1-style approaches, but that it can be more than an order of magnitude faster as well.

Data: For the low dimensional setting, the regressor $\mathbf{w}^* \in \mathbb{R}^p$ was chosen to be a random unit norm vector. Data was sampled as $\mathbf{x_i} \sim \mathcal{N}(0, I_p)$ and response variables were generated as $y_i^* = \langle \mathbf{w}^*, \mathbf{x}_i \rangle$. The set of corrupted points $\overline{S}_*$ was selected as a uniformly random (αn)-sized subset of $[n]$ and the corruptions were set to $b_i \sim U\left(-5\|\mathbf{y}^*\|_\infty, 5\|\mathbf{y}^*\|_\infty\right)$ for $i \in \overline{S}_*$. The corrupted responses were then generated as $y_i = y_i^* + b_i + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. For the sparse high-dimensional setting, $\text{supp}(\mathbf{w}^*)$ was selected to be a random s^*-sized subset of $[p]$. Phase-transition diagrams (Figure 1) were generated by repeating each experiment 100 times. For all other plots, each experiment was run over 20 random instances of the data and the plots were drawn to depict the mean results.

Algorithms: We compared various variants of our algorithm TORRENT to the regularized L_1 algorithm for robust regression [4, 5]. Note that the L_1 problem can be written as $\min_{\mathbf{z}} \|\mathbf{z}\|_1$ s.t.$A\mathbf{z} = \mathbf{y}$, where $A = \left[X^\top \; \frac{1}{\lambda} I_{m \times m}\right]$ and $\mathbf{z}^* = [\mathbf{w}^{*\top} \; \lambda \mathbf{b}^\top]^\top$. We used the Dual Augmented Lagrange Multiplier (DALM) L_1 solver implemented by [14] to solve the L_1 problem. We ran a fine tuned grid search over the λ parameter for the L_1 solver and quoted the best results obtained from the search. In the low-dimensional setting, we compared the recovery properties of TORRENT-FC (Algorithm 2) and TORRENT-HYB (Algorithm 4) with the DALM-L_1 solver, while for the high-dimensional case, we compared TORRENT-HD against the DALM-L_1 solver. Both the L_1 solver, as well as our methods, were implemented in Matlab and were run on a single core 2.4GHz machine with 8 GB RAM.

Choice of L_1-solver: An extensive comparative study of various L_1 minimization algorithms was performed by [14] who showed that the DALM and Homotopy solvers outperform other counterparts both in terms of recovery properties, and timings. We extended their study to our observation model and found the DALM solver to be significantly better than the other L_1 solvers; see Figure 3 in the appendix. We also observed, similar to [14], that the Approximate Message Passing (AMP) solver diverges on our problem as the input matrix to the L_1 solver is a non-Gaussian matrix $A = [X^T \frac{1}{\lambda} I]$.



(a) (b) (c) (d)

Figure 2: In low-dimensional (a,b), as well as sparse high dimensional (c,d) settings, TORRENT offers better recovery as the fraction of corrupted points α is varied. In terms of runtime, TORRENT is an order of magnitude faster than L_1 solvers in both settings. In the low-dim. setting, TORRENT-HYB is the fastest of all the variants.

Evaluation Metric: We measure the performance of various algorithms using the standard L_2 error: $r_{\widehat{\mathbf{w}}} = \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2$. For the phase-transition plots (Figure 1), we deemed an algorithm successful on an instance if it obtained a model $\widehat{\mathbf{w}}$ with error $r_{\widehat{\mathbf{w}}} < 10^{-4} \cdot \|\mathbf{w}^*\|_2$. We also measured the CPU time required by each of the methods, so as to compare their scalability.

6.1 Low Dimensional Results

Recovery Property: The phase-transition plots presented in Figure 1 represent our recovery experiments in graphical form. Both the fully-corrective and hybrid variants of TORRENT show better recovery properties than the L_1-minimization approach, indicated by the number of runs in which the algorithm was able to correctly recover $\mathbf{w}^*$ out of a 100 runs. Figure 2 shows the variation in recovery error as a function of α in the presence of white noise and exhibits the superiority of TORRENT-FC and TORRENT-HYB over L_1-DALM. Here again, TORRENT-FC and TORRENT-HYB achieve significantly lesser recovery error than L_1-DALM for all $\alpha <= 0.5$. Figure 3 in the appendix show that the variations of $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2$ with varying p, σ and n follow a similar trend with TORRENT having significantly lower recovery error in comparison to the L_1 approach.

Figure 1(d) brings out an interesting trend in the recovery property of TORRENT. As we increase the magnitude of corruption from $U\left(-\|\mathbf{y}^*\|_\infty, \|\mathbf{y}^*\|_\infty\right)$ to $U\left(-20\|\mathbf{y}^*\|_\infty, 20\|\mathbf{y}^*\|_\infty\right)$, the recovery error for TORRENT-HYB and TORRENT-FC decreases as expected since it becomes easier to identify the grossly corrupted points. However the L_1-solver was unable to exploit this observation and in fact exhibited an increase in recovery error.

Run Time: In order to ascertain the recovery guarantees for TORRENT on ill-conditioned problems, we performed an experiment where data was sampled as $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where $\text{diag}(\Sigma) \sim U(0, 5)$. Figure 2 plots the recovery error as a function of time. TORRENT-HYB was able to correctly recover $\mathbf{w}^*$ about $50\times$ faster than L_1-DALM which spent a considerable amount of time pre-processing the data matrix X. Even after allowing the L_1 algorithm to run for 500 iterations, it was unable to reach the desired residual error of 10^{-4}. Figure 2 also shows that our TORRENT-HYB algorithm is able to converge to the optimal solution much faster than TORRENT-FC or TORRENT-GD. This is because TORRENT-FC solves a least square problem at each step and thus, even though it requires significantly fewer iterations to converge, each iteration in itself is very expensive. While each iteration of TORRENT-GD is cheap, it is still limited by the slow $\mathcal{O}\left((1 - \frac{1}{\kappa})^t\right)$ convergence rate of the gradient descent algorithm, where κ is the condition number of the covariance matrix. TORRENT-HYB, on the other hand, is able to combine the strengths of both the methods to achieve faster convergence.

6.2 High Dimensional Results

Recovery Property: Figure 2 shows the variation in recovery error in the high-dimensional setting as the number of corrupted points was varied. For these experiments, n was set to $5s^* \log(p)$ and the fraction of corrupted points α was varied from 0.1 to 0.7. While L_1-DALM fails to recover $\mathbf{w}^*$ for $\alpha > 0.5$, TORRENT-HD offers perfect recovery even for α values upto 0.7.

Run Time: Figure 2 shows the variation in recovery error as a function of run time in this setting. L_1-DALM was found to be an order of magnitude slower than TORRENT-HD, making it infeasible for sparse high-dimensional settings. One key reason for this is that the L_1-DALM solver is significantly slower in identifying the set of clean points. For instance, whereas TORRENT-HD was able to identify the clean set of points in only 5 iterations, it took L_1 around 250 iterations to do the same.

References

[1] Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bölcskei. Recovery of Sparsely Corrupted Signals. *IEEE Transaction on Information Theory*, 58(5):3115–3130, 2012.

[2] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.

[3] John Wright, Alan Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[4] John Wright and Yi Ma. Dense Error Correction via ℓ^1 Minimization. *IEEE Transaction on Information Theory*, 56(7):3540–3560, 2010.

[5] Nam H. Nguyen and Trac D. Tran. Exact recoverability from dense corrupted observations via L1 minimization. *IEEE Transaction on Information Theory*, 59(4):2036–2058, 2013.

[6] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust Sparse Regression under Adversarial Corruption. In *30th International Conference on Machine Learning (ICML)*, 2013.

[7] Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M. Buhmann. Fast and Robust Least Squares Estimation in Corrupted Linear Models. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[8] Adrien-Marie Legendre (1805). On the Method of Least Squares. In (Translated from the French) D.E. Smith, editor, *A Source Book in Mathematics*, pages 576–579. New York: Dover Publications, 1959.

[9] Peter J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[10] Peter J. Rousseeuw and Katrien Driessen. Computing LTS Regression for Large Data Sets. *Journal of Data Mining and Knowledge Discovery*, 12(1):29–45, 2006.

[11] Thomas Blumensath and Mike E. Davies. Iterative Hard Thresholding for Compressed Sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[12] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On Iterative Hard Thresholding Methods for High-dimensional M-Estimation. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[13] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *26th International Conference on Machine Learning (ICML)*, 2009.

[14] Allen Y. Yang, Arvind Ganesh, Zihan Zhou, Shankar Sastry, and Yi Ma. A Review of Fast ℓ_1-Minimization Algorithms for Robust Face Recognition. CoRR abs/1007.3753, 2012.

[15] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[16] Thomas Blumensath. Sampling and reconstructing signals from a union of linear subspaces. *IEEE Transactions on Information Theory*, 57(7):4660–4671, 2011.

[17] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

A Convergence Guarantees with Dense Noise and Sparse Corruptions

We will now present recovery guarantees for the TORRENT-FC algorithm when both, dense noise, as well as sparse adversarial corruptions are present. Extensions for TORRENT-GD and TORRENT-HYB will follow similarly.

Theorem 10. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix and $\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b} + \boldsymbol{\varepsilon}$ be the corrupted output with sparse corruptions $\|\mathbf{b}\|_0 \leq \alpha \cdot n$ as well as dense bounded noise $\boldsymbol{\varepsilon}$. Let Algorithm 2 be executed on this data with the thresholding parameter set to $\beta \geq \alpha$. Let Σ_0 be an invertible matrix such that $\widetilde{X} = \Sigma_0^{-1/2} X$ satisfies the SSC and SSS properties at level γ with constants λ_γ and Λ_γ respectively (see Definition 1). If the data satisfies $\frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} < 1$, then after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 2 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon + C \frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{n}}$ for some constant $C > 0$.*

Proof. We being by observing that the optimality of the model $\mathbf{w}^{t+1}$ on the active set S_t ensures

$$\left\|\mathbf{y}_{S_t} - X_{S_t}^\top \mathbf{w}^{t+1}\right\|_2 = \left\|X_{S_t}^\top (\mathbf{w}^* - \mathbf{w}^{t+1}) + \boldsymbol{\varepsilon}_{S_t} + \mathbf{b}_{S_t}\right\|_2 \leq \left\|\mathbf{y}_t - X_{S_t}^\top \mathbf{w}^*\right\|_2 = \left\|\boldsymbol{\varepsilon}_{S_t} + \mathbf{b}_{S_t}\right\|_2,$$

which, upon the application of the triangle inequality, gives us

$$\left\|X_{S_t}^\top (\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2 \leq 2 \left\|\boldsymbol{\varepsilon}_{S_t} + \mathbf{b}_{S_t}\right\|_2.$$

Since $\left\|X_{S_t}^\top (\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2 \geq \sqrt{\lambda_{1-\beta}} \left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2$, we get

$$\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 \leq \frac{2}{\sqrt{\lambda_{1-\beta}}} \left\|\boldsymbol{\varepsilon}_{S_t} + \mathbf{b}_{S_t}\right\|_2 \leq \frac{2}{\sqrt{\lambda_{1-\beta}}} \left(\|\boldsymbol{\varepsilon}\|_2 + \|\mathbf{b}_{S_t}\|_2\right).$$

The hard thresholding step, on the other hand, guarantees that

$$\begin{aligned}
\left\|X_{S_{t+1}}^\top (\mathbf{w}^* - \mathbf{w}^{t+1}) + \boldsymbol{\varepsilon}_{S_{t+1}} + \mathbf{b}_{S_{t+1}}\right\|_2^2 &= \left\|\mathbf{y}_{S_{t+1}} - X_{S_{t+1}}^\top \mathbf{w}^{t+1}\right\|_2^2 \\
&\leq \left\|\mathbf{y}_{S_*} - X_{S_*}^\top \mathbf{w}^{t+1}\right\|_2 \\
&= \left\|X_{S_*}^\top (\mathbf{w}^* - \mathbf{w}^{t+1}) + \boldsymbol{\varepsilon}_{S_*}\right\|_2^2.
\end{aligned}$$

As before, let $\mathrm{CR}_{t+1} = S_{t+1} \backslash S_*$ and $\mathrm{MD}_{t+1} = S_* \backslash S_{t+1}$. Then we have

$$\left\|X_{\mathrm{CR}_{t+1}}^\top (\mathbf{w}^* - \mathbf{w}^{t+1}) + \boldsymbol{\varepsilon}_{\mathrm{CR}_{t+1}} + \mathbf{b}_{\mathrm{CR}_{t+1}}\right\|_2 \leq \left\|X_{\mathrm{MD}_{t+1}}^\top (\mathbf{w}^* - \mathbf{w}^{t+1}) + \boldsymbol{\varepsilon}_{\mathrm{MD}_{t+1}}\right\|_2.$$

An application of the triangle inequality and the fact that $\left\|\mathbf{b}_{\mathrm{CR}_{t+1}}\right\|_2 = \left\|\mathbf{b}_{S_{t+1}}\right\|$ gives us

$$\begin{aligned}
\left\|\mathbf{b}_{S_{t+1}}\right\|_2 &\leq \left\|X_{\mathrm{MD}_{t+1}}^\top (\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2 + \left\|X_{\mathrm{CR}_{t+1}}^\top (\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2 + \left\|\boldsymbol{\varepsilon}_{\mathrm{CR}_{t+1}}\right\|_2 + \left\|\boldsymbol{\varepsilon}_{\mathrm{MD}_{t+1}}\right\|_2 \\
&\leq 2\sqrt{\Lambda_\beta} \left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 + \sqrt{2} \|\boldsymbol{\varepsilon}\|_2, \\
&= \frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} \|\mathbf{b}_{S_t}\|_2 + \left(\frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} + \sqrt{2}\right) \|\boldsymbol{\varepsilon}\|_2 \\
&\leq \eta \cdot \|\mathbf{b}_{S_t}\|_2 + (1 + \sqrt{2}) \|\boldsymbol{\varepsilon}\|_2,
\end{aligned}$$

where the second step uses the fact that $\max\left\{|\mathrm{CR}_{t+1}|, |\mathrm{MD}_{t+1}|\right\} \leq \beta \cdot n$ and the Cauchy-Schwartz inequality, and the last step uses the fact that for sufficiently small β, we have $\eta := \frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}}$. Using the inequality for $\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2$ again gives us

$$\begin{aligned}
\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 &\leq \frac{2}{\sqrt{\lambda_{1-\beta}}} \left(\|\boldsymbol{\varepsilon}\|_2 + \|\mathbf{b}_{S_t}\|_2\right) \\
&\leq \frac{4 + 2\sqrt{2}}{\sqrt{\lambda_{1-\beta}}} \|\boldsymbol{\varepsilon}\|_2 + \frac{2 \cdot \eta^t}{\sqrt{\lambda_{1-\beta}}} \|\mathbf{b}\|_2
\end{aligned}$$

For large enough n we have $\sqrt{\lambda_{1-\beta}} \geq \mathcal{O}(\sqrt{n})$, which completes the proof. $\square$

Notice that for random Gaussian noise, this result gives the following convergence guarantee.

Corollary 11. *Let the date be generated as before with random Gaussian dense noise i.e.* $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b} + \boldsymbol{\varepsilon}$ *with* $\|\mathbf{b}\|_0 \leq \alpha \cdot n$ *and* $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I)$*. Let Algorithm 2 be executed on this data with the thresholding parameter set to* $\beta \geq \alpha$*. Let* Σ_0 *be an invertible matrix such that* $\widetilde{X} = \Sigma_0^{-1/2}X$ *satisfies the SSC and SSS properties at level* γ *with constants* λ_γ *and* Λ_γ *respectively (see Definition 1). If the data satisfies* $\frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} < 1$*, then after* $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ *iterations, Algorithm 2 obtains an* ϵ*-accurate solution* $\mathbf{w}^t$ *i.e.* $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon + 2\sigma C$*, where* $C > 0$ *is the constant in Theorem 10.*

Proof. Using tail bounds on Chi-squared distributions [15], we get, with probability at least $1 - \delta$,

$$\|\boldsymbol{\varepsilon}\|_2^2 \leq \sigma^2\left(n + 2\sqrt{n\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}\right).$$

Thus, for $n > 4\log\frac{1}{\delta}$, we have $\|\boldsymbol{\varepsilon}\|_2^2 \leq 2\sigma n$ which proves the result. □

Remark 6. We note that the design assumptions made by Theorem 10 (i..e $\frac{4\sqrt{\Lambda_\beta}}{\sqrt{\lambda_{1-\beta}}} < 1$) are similar to those made by Theorem 3 and would be satisfied with high probability by data sampled from sub-Gaussian distributions (see Appendix G for details).

B Proof of Theorem 3

Theorem 3. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ be the given data matrix and $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}$ be the corrupted output with $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Let Algorithm 2 be executed on this data with the thresholding parameter set to $\beta \geq \alpha$. Let Σ_0 be an invertible matrix such that $\widetilde{X} = \Sigma_0^{-1/2}X$ satisfies the SSC and SSS properties at level γ with constants λ_γ and Λ_γ respectively (see Definition 1). If the data satisfies $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < 1$, then after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 2 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$.

Proof. Let $\mathbf{r}^t = \mathbf{y} - X^\top\mathbf{w}^t$ be the vector of residuals at time t and $C_t = X_{S_t}X_{S_t}^\top$. Since $\lambda_\alpha > 0$ (something which we shall establish later), we get

$$\mathbf{w}^{t+1} = C_t^{-1}X_{S_t}\mathbf{y}_{S_t} = C_t^{-1}X_{S_t}\left(X_{S_t}^\top\mathbf{w}^* + \mathbf{b}_{S_t}\right) = \mathbf{w}^* + C_t^{-1}X_{S_t}\mathbf{b}_{S_t}.$$

Thus, for any set $S \subset [n]$, we have

$$\mathbf{r}_S^{t+1} = \mathbf{y}_S - X_S^\top\mathbf{w}_{t+1} = \mathbf{b}_S - X_S^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}$$

This, gives us

$$\begin{aligned}
\|\mathbf{b}_{S_{t+1}}\|_2^2 &= \left\|\mathbf{b}_{S_{t+1}} - X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 - \left\|X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 + 2\cdot\mathbf{b}_{S_{t+1}}^\top X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t} \\
&\overset{\zeta_1}{\leq} \left\|\mathbf{b}_{S_*} - X_{S_*}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 - \left\|X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 + 2\cdot\mathbf{b}_{S_{t+1}}^\top X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t} \\
&\overset{\zeta_2}{=} \left\|X_{S_*}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 - \left\|X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 + 2\cdot\mathbf{b}_{S_{t+1}}^\top X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t} \\
&\leq \left\|X_{S_*\backslash S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t}\right\|_2^2 + 2\cdot\mathbf{b}_{S_{t+1}}^\top X_{S_{t+1}}^\top C_t^{-1}X_{S_t}\mathbf{b}_{S_t} \\
&\overset{\zeta_3}{=} \left\|\widetilde{X}_{S_*\backslash S_{t+1}}^\top\left(\widetilde{X}_{S_t}\widetilde{X}_{S_t}^T\right)^{-1}\widetilde{X}_{S_t}\mathbf{b}_{S_t}\right\|_2^2 + 2\cdot\mathbf{b}_{S_{t+1}}^\top\widetilde{X}_{S_{t+1}}^\top\left(\widetilde{X}_{S_t}\widetilde{X}_{S_t}^T\right)^{-1}\widetilde{X}_{S_t}\mathbf{b}_{S_t} \\
&\overset{\zeta_4}{\leq} \frac{\Lambda_\beta^2}{\lambda_{1-\beta}^2}\cdot\|\mathbf{b}_{S_t}\|_2^2 + 2\cdot\frac{\Lambda_\beta}{\lambda_{1-\beta}}\cdot\|\mathbf{b}_{S_t}\|_2\left\|\mathbf{b}_{S_{t+1}}\right\|_2,
\end{aligned}$$

where ζ_1 follows since the hard thresholding step ensures $\left\|\mathbf{r}^{t+1}_{S_{t+1}}\right\|_2^2 \leq \left\|\mathbf{r}^{t+1}_{S_*}\right\|_2^2$ (see Claim 19 and use the fact that $\beta \geq \alpha$), ζ_2 notices the fact that $\mathbf{b}_{S_*} = \mathbf{0}$. ζ_3 follows from setting $\widetilde{X} = \Sigma_0^{-1/2} X$ and $X_S^\top C_t^{-1} X_{S'} = \widetilde{X}_S^\top (\widetilde{X}_{S_t} \widetilde{X}_{S_t}^\top)^{-1} \widetilde{X}_{S'}$. ζ_4 follows from the definition of SSC and SSS properties, $\|\mathbf{b}_{S_t}\|_0 \leq \|\mathbf{b}\|_0 \leq \beta \cdot n$ and $|S_* \backslash S_{t+1}| \leq \beta \cdot n$. Solving the quadratic equation gives us

$$\left\|\mathbf{b}_{S_{t+1}}\right\|_2 \leq (1+\sqrt{2}) \cdot \frac{\Lambda_\beta}{\lambda_{1-\beta}} \cdot \|\mathbf{b}_{S_t}\|_2 . \tag{4}$$

Let $\eta := \frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}}$ denote the convergence rate in (4). We shall show below that for a large family of random designs, we have $\eta < 1$ if $n \geq \Omega\left(p + \log \frac{1}{\delta}\right)$. We now recall from our earlier discussion that $\mathbf{w}^{t+1} = \mathbf{w}^* + C_t^{-1} X_{S_t} \mathbf{b}_{S_t}$ which gives us

$$\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 = \left\|C_t^{-1} X_{S_t} \mathbf{b}_{S_t}\right\|_2 \leq \frac{\sqrt{\Lambda_\beta}}{\lambda_{1-\beta}} \cdot \|\mathbf{b}_{S_t}\|_2 \leq \eta^t \cdot \frac{\sqrt{\Lambda_\beta}}{\lambda_{1-\beta}} \|\mathbf{b}\|_2 \leq \epsilon,$$

for $t \geq \log_{\frac{1}{\eta}} \left(\frac{\sqrt{\Lambda_\beta}}{\lambda_{1-\beta}} \cdot \frac{\|\mathbf{b}\|_2}{\epsilon} \right)$. Noting that $\frac{\sqrt{\Lambda_\beta}}{\lambda_{1-\beta}} \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ establishes the convergence result. □

C Proof of Theorem 4

Theorem 4. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix with each $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Let $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$ and $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Also, let $\alpha \leq \beta < \frac{1}{65}$ and $n \geq \Omega\left(p + \log \frac{1}{\delta}\right)$. Then, with probability at least $1-\delta$, the data satisfies $\frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < \frac{9}{10}$. More specifically, after $T \geq 10 \log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)$ iterations of Algorithm 1 with the thresholding parameter set to β, we have $\left\|\mathbf{w}^T - \mathbf{w}^*\right\| \leq \epsilon$.

Proof. We note that whenever $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ then $\Sigma^{-1/2}\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$. Thus, Theorem 15 assures us that with probability at least $1 - \delta$, the data matrix $\widetilde{X} = \Sigma^{-1/2} X$ satisfies the SSC and SSS properties with the following constants

$$\Lambda_\beta \leq \beta n \left(1 + 3e\sqrt{6 \log \frac{e}{\beta}}\right) + \mathcal{O}\left(\sqrt{np + n \log \frac{1}{\delta}}\right)$$

$$\lambda_{1-\beta} \geq n - \beta n \left(1 + 3e\sqrt{6 \log \frac{e}{\beta}}\right) - \Omega\left(\sqrt{np + n \log \frac{1}{\delta}}\right)$$

Thus, the convergence given be Algorithm 1, when invoked with $\Sigma_0 = \Sigma$, relies on the quantity $\eta = \frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}}$ being less than unity. This translates to the requirement $(1+\sqrt{2})\Lambda_\beta \leq \lambda_{1-\beta}$. Using the above bounds translates that requirement to

$$\underbrace{(2+\sqrt{2})\beta \left(1 + 3e\sqrt{6 \log \frac{e}{\beta}}\right)}_{(A)} + \underbrace{\mathcal{O}\left(\sqrt{\frac{p}{n} + \frac{1}{n} \log \frac{1}{\delta}}\right)}_{(B)} < 1.$$

For $n = \Omega\left(p + \log \frac{1}{\delta}\right)$, the second quantity (B) can be made as small a constant as necessary. Tackling the first quantity (A) turns out to be more challenging. However, we can show that for all $\beta < \frac{1}{190}$, we get $\eta = \frac{(1+\sqrt{2})\Lambda_\beta}{\lambda_{1-\beta}} < \frac{9}{10}$ which establishes the claimed result. Thus, Algorithm 1 can tolerate a corruption index of upto $\alpha \leq \frac{1}{190}$. However, we note that using a more finely tuned setting of the constant ϵ in the proof of Theorem 15 and a more careful proof using tight tail inequalities for chi-squared distributions [15], we can achieve a better corruption level tolerance of $\alpha < \frac{1}{65}$. □

D Proof of Theorem 5

Theorem 5. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix and $\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b}$ be the corrupted output with $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Let X satisfy the SSC and SSS properties at level γ with constants λ_γ and Λ_γ respectively (see Definition 1). Let Algorithm 1 be executed on this data with the GD update (Algorithm 3) with the thresholding parameter set to $\beta \geq \alpha$ and the step length set to $\eta = \frac{1}{\Lambda_{1-\beta}}$. If the data satisfies $\max\left\{\eta\sqrt{\Lambda_\beta}, 1 - \eta\lambda_{1-\beta}\right\} \leq \frac{1}{4}$, then after $t = \mathcal{O}\left(\log\left(\frac{\|b\|_2}{\sqrt{n}}\frac{1}{\epsilon}\right)\right)$ iterations, Algorithm 1 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq \epsilon$.

Proof. Let $\mathbf{r}^t = \mathbf{y} - X^\top \mathbf{w}^t$ be the vector of residuals at time t and $C_t = X_{S_t} X_{S_t}^\top$. We have

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \eta \cdot X_{S_t} \mathbf{r}^t_{S_t} = \mathbf{w}^t + \eta \cdot X_{S_t}(\mathbf{y}_{S_t} - X_{S_t}^\top \mathbf{w}^t)$$

The thresholding step ensures that $\left\|\mathbf{r}^{t+1}_{S_{t+1}}\right\|_2^2 \leq \left\|\mathbf{r}^{t+1}_{S_*}\right\|_2^2$ (see Claim 19 and use $\beta \geq \alpha$) which implies

$$\left\|\mathbf{r}^{t+1}_{\mathrm{CR}_{t+1}}\right\|_2^2 \leq \left\|\mathbf{r}^{t+1}_{\mathrm{MD}_{t+1}}\right\|_2^2,$$

where $\mathrm{CR}_{t+1} = S_{t+1} \backslash S_*$ are the *corrupted recoveries* and $\mathrm{MD}_{t+1} = S_* \backslash S_{t+1}$ are the clean points *missed* out from *detection*. Note that $|\mathrm{CR}_{t+1}| \leq \alpha \cdot n$ and $|\mathrm{MD}_{t+1}| \leq \beta \cdot n$. Since $\mathbf{b}_{S_*} = \mathbf{0}$ and $\mathrm{MD}_{t+1} \subseteq S_*$, we get

$$\left\|\mathbf{b}_{\mathrm{CR}_{t+1}} + X^\top_{\mathrm{CR}_{t+1}}(\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2 \leq \left\|X^\top_{\mathrm{MD}_{t+1}}(\mathbf{w}^* - \mathbf{w}^{t+1})\right\|_2$$

Using the SSS conditions and the fact that $\left\|\mathbf{b}_{S_{t+1}}\right\|_2 = \left\|\mathbf{b}_{S_{t+1}\backslash S_*}\right\|_2$ gives us

$$\left\|\mathbf{b}_{S_{t+1}}\right\|_2 = \left\|\mathbf{b}_{\mathrm{CR}_{t+1}}\right\|_2 \leq (\sqrt{\Lambda_\alpha} + \sqrt{\Lambda_\beta})\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 \leq 2\sqrt{\Lambda_\beta}\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2$$

Now, using the expression for $\mathbf{w}^{t+1}$ gives us

$$\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 \leq \left\|(I - \eta C_t)(\mathbf{w}^* - \mathbf{w}^t)\right\|_2 + \eta\left\|X_{S_t}\mathbf{b}_{S_t}\right\|_2$$

We will bound the two terms on the right hand separately. We can bound the second term easily as

$$\eta\left\|X_{S_t}\mathbf{b}_{S_t}\right\|_2 \leq \eta\sqrt{\Lambda_\alpha}\left\|\mathbf{b}_{S_t}\right\|_2 \leq \eta\sqrt{\Lambda_\beta}\left\|\mathbf{b}_{S_t}\right\|_2,$$

since $\left\|\mathbf{b}_{S_t}\right\|_0 \leq \alpha \cdot n$. For the first term we observe that for $\eta \leq \frac{1}{\Lambda_{1-\beta}}$, we have

$$\left\|I - \eta C_t\right\|_2 = \sup_{\mathbf{v} \in S^{p-1}}\left|1 - \eta \cdot \mathbf{v}^\top C_t \mathbf{v}\right| = \sup_{\mathbf{v} \in S^{p-1}}\left\{1 - \eta \cdot \mathbf{v}^\top C_t \mathbf{v}\right\} \leq 1 - \eta\lambda_{1-\beta},$$

which we can use to bound

$$\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 \leq (1 - \eta\lambda_{1-\beta})\left\|\mathbf{w}^* - \mathbf{w}^t\right\|_2 + \eta\sqrt{\Lambda_\beta}\left\|\mathbf{b}_{S_t}\right\|_2$$

This gives us, for $\eta = \frac{1}{\Lambda_{1-\beta}}$,

$$\left\|\mathbf{b}_{S_{t+1}}\right\|_2 \leq 2\sqrt{\Lambda_\beta}\left\|\mathbf{w}^* - \mathbf{w}^{t+1}\right\|_2 \leq 2\underbrace{\left(1 - \frac{\lambda_{1-\beta}}{\Lambda_{1-\beta}}\right)}_{(P)}\sqrt{\Lambda_\beta}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|_2 + 2\underbrace{\frac{\Lambda_\beta}{\Lambda_{1-\beta}}}_{(Q)}\left\|\mathbf{b}_{S_t}\right\|_2.$$

For Gaussian designs and small enough β, we can show $(Q) \leq \frac{1}{4}$ as we did in Theorem 4. To bound (P), we use the lower bound on $\lambda_{1-\beta}$ given by Theorem 15 and use the following tighter upper bound for $\Lambda_{1-\beta}$:

$$\Lambda_{1-\beta} \leq \left((1-\beta) + 3e\sqrt{6\beta(1-\beta)\log\frac{e}{\beta}}\right)n + \mathcal{O}\left(\sqrt{np + n\log\frac{1}{\delta}}\right)$$

The above bound is obtained similarly to the one in Theorem 15 but uses the identity $\binom{n}{k} = \binom{n}{n-k} \leq \left(\frac{en}{n-k}\right)^{n-k}$ for values of $k \geq n/2$ instead. For small enough β and $n = \Omega\left(\kappa^2(\Sigma)(p + \log\frac{1}{\delta})\right)$, we can then show $(P) \leq \frac{1}{4}$ as well. Let $\Psi_t := \sqrt{n}\left\|\mathbf{w}^* - \mathbf{w}^t\right\|_2 + \left\|b_{S_t}\right\|$. Using elementary manipulations and the fact that $\sqrt{\Lambda_\beta} \geq \Omega\left(\sqrt{n}\right)$, we can then show that

$$\Psi_{t+1} \leq 3/4 \cdot \Psi_t.$$

Thus, in $t = \mathcal{O}\left(\log\left(\left(\left\|\mathbf{w}^*\right\|_2 + \frac{\|b\|_2}{\sqrt{n}}\right)\frac{1}{\epsilon}\right)\right)$ iterations of the algorithm, we arrive at an ϵ-optimal solution i.e. $\left\|\mathbf{w}^* - \mathbf{w}^t\right\|_2 \leq \epsilon$. A similar argument holds true for sub-Gaussian designs as well. $\square$

E Proof of Theorem 6

Theorem 6. Suppose Algorithm 4 is executed on data that allows Algorithms 2 and 3 a convergence rate of η_{FC} and η_{GD} respectively. Suppose we have $2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} < 1$. Then for *any* interleavings of the FC and GD steps that the policy may enforce, after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 4 ensures an ϵ-optimal solution i.e. $\|\mathbf{w}^t - \mathbf{w}^*\| \leq \epsilon$.

Proof. Our proof shall essentially show that the FC and GD steps do not undo the progress made by the other if executed in succession and if $2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} < 1$, actually ensure non-trivial progress. Let

$$\Psi_t^{\text{FC}} = \|\mathbf{b}_{S_t}\|_2$$
$$\Psi_t^{\text{GD}} = \sqrt{n}\left\|\mathbf{w}^t - \mathbf{w}^*\right\| + \|\mathbf{b}_{S_t}\|_2$$

denote the potential functions used in the analyses of the FC and GD algorithms before. Then we will show below that if the FC and GD algorithms are executed in steps t and $t+1$ then we have

$$\Psi_{t+2}^{\text{FC}} \leq 2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} \cdot \Psi_t^{\text{FC}}$$

Alternatively, if the GD and FC algorithms are executed in steps t and $t+1$ respectively, then

$$\Psi_{t+2}^{\text{GD}} \leq 2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} \cdot \Psi_t^{\text{GD}}$$

Thus, if algorithm executes the FC step at the time step t, then it would at least ensure $\Psi_t^{\text{FC}} \leq (2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}})^{t/2} \cdot \Psi_0^{\text{FC}}$ (similarly if the last step is a GD step). Since both the FC and GD algorithms ensure $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$ for $t \geq \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}} \frac{\|b\|_2}{\epsilon}\right)\right)$, the claim would follow.

We now prove the two claimed results regarding the two types of interleaving below

1. FC $\longrightarrow$ GD
   The FC step guarantees $\left\|\mathbf{b}_{S_{t+1}}\right\|_2 \leq \eta_{\text{FC}} \cdot \|\mathbf{b}_{S_t}\|$ as well as $\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 \leq \eta_{\text{FC}} \cdot \frac{\|\mathbf{b}_{S_t}\|}{\sqrt{n}}$, whereas the GD step guarantees $\Psi_{t+2}^{\text{GD}} \leq \eta_{\text{GD}} \cdot \Psi_{t+1}^{\text{GD}}$. Together these guarantee
   $$\begin{aligned}\sqrt{n}\left\|\mathbf{w}^{t+2} - \mathbf{w}^*\right\|_2 + \left\|\mathbf{b}_{S_{t+2}}\right\|_2 &\leq \eta_{\text{GD}} \cdot \sqrt{n}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 + \left\|\mathbf{b}_{S_{t+1}}\right\|_2 \\ &\leq 2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} \cdot \|\mathbf{b}_{S_t}\|_2\end{aligned}$$
   Since $\sqrt{n}\left\|\mathbf{w}^{t+2} - \mathbf{w}^*\right\|_2 \geq 0$, this yields the result.
2. GD $\longrightarrow$ FC
   The GD step guarantees $\Psi_{t+1}^{\text{GD}} \leq \eta_{\text{GD}} \cdot \Psi_t^{\text{GD}}$ whereas the FC step guarantees $\left\|\mathbf{b}_{S_{t+2}}\right\|_2 \leq \eta_{\text{FC}} \cdot \left\|\mathbf{b}_{S_{t+1}}\right\|$ as well as $\left\|\mathbf{w}^{t+2} - \mathbf{w}^*\right\|_2 \leq \eta_{\text{FC}} \cdot \frac{\left\|\mathbf{b}_{S_{t+1}}\right\|}{\sqrt{n}}$. Together these guarantee
   $$\begin{aligned}\sqrt{n}\left\|\mathbf{w}^{t+2} - \mathbf{w}^*\right\|_2 + \left\|\mathbf{b}_{S_{t+2}}\right\|_2 &\leq 2\eta_{\text{FC}}\left\|\mathbf{b}_{S_{t+1}}\right\|_2 \\ &\leq 2 \cdot \eta_{\text{FC}} \cdot \eta_{\text{GD}} \cdot \Psi_t^{\text{GD}},\end{aligned}$$
   where the second step follows from the GD step guarantee since $\sqrt{n}\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 \geq 0$.

This finishes the proof. □

F Proof of Theorem 9

Theorem 9. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ be the given data matrix and $\mathbf{y} = X^T\mathbf{w}^* + \mathbf{b}$ be the corrupted output with $\|\mathbf{w}^*\|_0 \leq s^*$ and $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Let Algorithm 2 be executed on this data with the IHT update from [12] and thresholding parameter set to $\beta \geq \alpha$. Let Σ_0 be an invertible matrix such that $\Sigma_0^{-1/2}X$ satisfies the SRSC and SRSS properties at level $(\gamma, 2s+s^*)$ with constants $\alpha_{(\gamma,2s+s^*)}$ and $L_{(\gamma,2s+s^*)}$ respectively (see Definition 8) for $s \geq 32\left(\frac{L_{(\gamma,2s+s^*)}}{\alpha_{(\gamma,2s+s^*)}}\right)$ with $\gamma = 1-\beta$. If X also satisfies $\frac{4L_{(\beta,s+s^*)}}{\alpha_{(1-\beta,s+s^*)}} < 1$, then after $t = \mathcal{O}\left(\log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$ iterations, Algorithm 2 obtains an ϵ-accurate solution $\mathbf{w}^t$ i.e. $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$. In particular, if X is sampled from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ and $n \geq \Omega\left((2s + s^*)\log p + \log \frac{1}{\delta}\right)$, then for all values of $\alpha \leq \beta < \frac{1}{65}$, we can guarantee recovery as $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$.

Proof. We first begin with the guarantee provided by existing sparse recovery techniques. The results of [12], for example, indicate that if the input to the algorithm indeed satisfies the RSC and RSS properties at the level $(1-\beta, 2s+s^*)$ with constants α_{2s+s^*} and L_{2s+s^*} for $s \geq 32\left(\frac{L_{2s+s^*}}{\alpha_{2s+s^*}}\right)$, then in time $\tau = \mathcal{O}\left(\frac{L_{2s+s^*}}{\alpha_{2s+s^*}} \cdot \log\left(\frac{\|b\|_2}{\rho}\right)\right)$, the IHT algorithm [12, Algorithm 1] outputs an updated model $\mathbf{w}^{t+1}$ that satisfies $\left\|\mathbf{w}^{t+1}\right\|_0 \leq s$, as well as

$$\left\|\mathbf{y}_{S_t} - X_{S_t}^\top \mathbf{w}^{t+1}\right\|_2^2 \leq \left\|\mathbf{y}_{S_t} - X_{S_t}^\top \mathbf{w}^*\right\|_2^2 + \rho.$$

We will set ρ later. Since the SRSC and SRSS properties ensure the above and $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$, this gives us

$$\left\|X_{S_t}^\top(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2^2 \leq 2(\mathbf{w}^{t+1} - \mathbf{w}^*)^\top X_{S_t}^\top \mathbf{b}_{S_t} + \rho = 2(\mathbf{w}^{t+1} - \mathbf{w}^*)^\top X_{S_t \cap \bar{S}_*}^\top \mathbf{b}_{S_t \cap \bar{S}_*} + \rho,$$

since $\mathbf{b}_S = \mathbf{0}$ for any set $S \cap \bar{S}_* = \phi$. We now analyze the two sides separately below using the SRSC and SRSS properties below. For any $S \subset [n]$, denote $\tilde{X}_S := \Sigma_0^{-1/2} X$.

$$\left\|X_{S_t}^\top(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2^2 = \left\|\tilde{X}_{S_t}^\top \Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2^2 \geq \alpha_{(1-\beta, s+s^*)} \left\|\Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2^2$$

$$\left\|X_{S_t \cap \bar{S}_*}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\| = \left\|\tilde{X}_{S_t \cap \bar{S}_*} \Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\| \leq \sqrt{L_{(\beta, s+s^*)}} \left\|\Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2.$$

Now, if $\left\|\mathbf{w}^{t+1} - \mathbf{w}^*\right\|_2 \geq \epsilon$, then $\left\|\Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 \geq \sqrt{\lambda_{\min}(\Sigma_0)} \cdot \epsilon$. This give us

$$\begin{aligned}
\left\|\Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 &\leq \frac{2\sqrt{L_{(\beta, s+s^*)}}}{\alpha_{(1-\beta, s+s^*)}} \left\|\mathbf{b}_{S_t \cap \bar{S}_*}\right\|_2 + \frac{\rho}{\alpha_{(1-\beta, s+s^*)}} \\
&= \frac{2\sqrt{L_{(\beta, s+s^*)}}}{\alpha_{(1-\beta, s+s^*)}} \left\|\mathbf{b}_{S_t}\right\|_2 + \frac{\rho}{\epsilon \cdot \sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta, s+s^*)}}.
\end{aligned}$$

We note that although we declared the SRSC and SRSS properties for the action of matrices on sparse vectors (such as $\mathbf{w}^* - \mathbf{w}^{t+1}$), we instead applied them above to the action of matrices on sparse vectors transformed by $\Sigma_0^{1/2}$ ($\Sigma_0^{1/2}(\mathbf{w}^* - \mathbf{w}^{t+1})$). Since $\Sigma_0^{1/2}\mathbf{v}$ need not be sparse even if $\mathbf{v}$ is sparse, this appears to pose a problem. However, all we need to resolve this is to notice that the proof technique of Theorem 18 which would be used to establish the SRSC and SRSS properties, holds in general for not just the action of a matrix on the set of sparse vectors, but on vectors in the union of any fixed set of low dimensional subspaces.

More specifically, we can modify the RSC and RSS properties (and by extension, the SRSC and SRSS properties), to requiring that the matrix X act as an approximate isometry on the following set of vectors $S_{(s,\Sigma_0)}^{p-1} := \left\{\mathbf{v} : \mathbf{v} = \Sigma_0^{-1/2}\mathbf{v}' \text{ for some } \mathbf{v}' \in S_s^{p-1}\right\}$. We refer the reader to the work of [16] which describes this technique in great detail. Proceeding with the proof, the assurance of the thresholding step, as used in the proof of Theorem 5, along with a straightforward application of the (modified) SRSS property gives us

$$\begin{aligned}
\left\|\mathbf{b}_{S_{t+1}}\right\|_2 &\leq \left\|X_{\mathrm{CR}_{t+1}}^\top(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 + \left\|X_{\mathrm{MD}_{t+1}}^\top(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 \\
&= \left\|\tilde{X}_{\mathrm{CR}_{t+1}}^\top \Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 + \left\|\tilde{X}_{\mathrm{MD}_{t+1}}^\top \Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 \\
&\leq 2\sqrt{L_{(\beta, s+s^*)}} \left\|\Sigma_0^{1/2}(\mathbf{w}^{t+1} - \mathbf{w}^*)\right\|_2 \\
&\leq \frac{4L_{(\beta, s+s^*)}}{\alpha_{(1-\beta, s+s^*)}} \left\|\mathbf{b}_{S_t}\right\|_2 + \frac{2\rho\sqrt{L_{(\beta, s+s^*)}}}{\epsilon \cdot \sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta, s+s^*)}}
\end{aligned}$$

Thus, whenever $\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|_2 > \epsilon$, in successive steps, $\|\mathbf{b}_{S_t}\|_2$ undergoes a linear decrease. Denoting $\eta := \frac{4L_{(\beta,s+s^*)}}{\alpha_{(1-\beta,s+s^*)}}$, we get

$$\left\|\mathbf{b}_{S_{t+1}}\right\|_2 \leq \eta^t \cdot \|\mathbf{b}\|_2 + \left(\frac{1-\eta^t}{1-\eta}\right) \frac{2\rho\sqrt{L_{(\beta,s+s^*)}}}{\epsilon \cdot \sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta,s+s^*)}}$$

and using $\left\|\Sigma_0^{1/2}(\mathbf{w}^t-\mathbf{w}^*)\right\|_2 \geq \sqrt{\lambda_{\min}(\Sigma_0)}\,\|\mathbf{w}^t-\mathbf{w}^*\|_2$ gives us

$$\begin{aligned}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|_2 &\leq \frac{2\sqrt{L_{(\beta,s+s^*)}}}{\sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta,s+s^*)}} \left\|\mathbf{b}_{S_{t+1}}\right\|_2 + \frac{\rho}{\lambda_{\min}(\Sigma_0) \cdot \alpha_{(1-\beta,s+s^*)}} \\ &\leq \eta^t \frac{2\sqrt{L_{(\beta,s+s^*)}}}{\sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta,s+s^*)}} \|\mathbf{b}\|_2 + \frac{36\rho}{\epsilon \cdot \lambda_{\min}(\Sigma_0) \cdot \alpha_{(1-\beta,s+s^*)}},\end{aligned}$$

where we have assumed that $\frac{4L_{(\beta,s+s^*)}}{\alpha_{(1-\beta,s+s^*)}} < 9/10$, something that we shall establish below. Note that $\lambda_{\min}(\Sigma_0) > 0$ since Σ is assumed to be invertible. In the random design settings we shall consider, we also have $\frac{\sqrt{L_{(\beta,s+s^*)}}}{\sqrt{\lambda_{\min}(\Sigma_0)} \cdot \alpha_{(1-\beta,s+s^*)}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Then setting $\rho \leq \frac{1}{72}\epsilon^2 \cdot \lambda_{\min}(\Sigma_0) \cdot \alpha_{(1-\beta,s+s^*)}$ proves the convergence result.

As before, we can use the above result to establish sparse recovery guarantees in the statistical setting for Gaussian and sub-Gaussian design models. If our data matrix X is generated from a Gaussian distribution $\mathcal{N}(\mathbf{0},\Sigma)$ for some invertible Σ, then the results in Theorem 18 can be used to establish that $\Sigma^{-1/2}X$ satisfies the SRSC and SRSS properties at the required levels and that for $\alpha < \frac{1}{190}$ and $n \geq \Omega\left((2s+s^*)\log p + \log\frac{1}{\delta}\right)$, we have $\eta = \frac{2L_{(\beta,s+s^*)}}{\alpha_{(1-\beta,s+s^*)}} < 9/10$.

Thus, the above result can be applied with $\Sigma_0 = \Sigma$ to get convergence guarantees in the general Gaussian setting. We note that the above analysis can tolerate the same level of corruption as Theorem 4 and thus, we can improve the noise tolerance level to $\alpha \leq \frac{1}{65}$ here as well. We also note that these results can be readily extended to the sub-Gaussian setting as well. $\square$

G Robust Statistical Estimation

This section elaborates on how results on the convergence guarantees of our algorithms can be used to give guarantees for robust statistical estimation problems. We begin with a few definition of sampling models that would be used in our results.

Definition 12. *A random variable $x \in \mathbb{R}$ is called sub-Gaussian if the following quantity is finite*

$$\sup_{p\geq 1} p^{-1/2}\left(\mathbb{E}\,|x|^p\right)^{1/p}.$$

Moreover, the smallest upper bound on this quantity is referred to as the sub-Gaussian norm of x and denoted as $\|x\|_{\psi_2}$.

Definition 13. *A vector-valued random variable $\mathbf{x} \in \mathbb{R}^p$ is called sub-Gaussian if its unidimensional marginals $\langle \mathbf{x},\mathbf{v}\rangle$ are sub-Gaussian for all $\mathbf{v} \in S^{p-1}$. Moreover, its sub-Gaussian norm is defined as follows*

$$\|X\|_{\psi_2} := \sup_{\mathbf{v}\in S^{p-1}} \|\langle \mathbf{x},\mathbf{v}\rangle\|_{\psi_2}$$

We will begin with the analysis of Gaussian designs and then extend our analysis for the class of general sub-Gaussian designs.

Lemma 14. *Let $X \in \mathbb{R}^{p\times n}$ be a matrix whose columns are sampled i.i.d from a standard Gaussian distribution i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},I)$. Then for any $\epsilon > 0$, with probability at least $1-\delta$, X satisfies*

$$s_{\max}(XX^\top) \leq n + (1-2\epsilon)^{-1}\sqrt{cnp + c'n\log\frac{2}{\delta}}$$

$$s_{\min}(XX^\top) \geq n - (1-2\epsilon)^{-1}\sqrt{cnp + c'n\log\frac{2}{\delta}},$$

where $c = 24e^2\log\frac{3}{\epsilon}$ and $c' = 24e^2$.

Proof. We will first use the fact that X is sampled from a standard Gaussian to show that its covariance concentrates around identity. Thus, we first show that with high probability,

$$\left\|XX^\top - nI\right\|_2 \leq \epsilon_1$$

for some $\epsilon_1 < 1$. Doing so will automatically establish the following result

$$n - \epsilon_1 \leq s_{\min}(XX^\top) \leq s_{\max}(XX^\top) \leq n + \epsilon_1.$$

Let $A := XX^\top - I$. We will use the technique of covering numbers [17] to establish the above. Let $\mathcal{C}^{p-1}(\epsilon) \subset S^{p-1}$ be an ϵ cover for S^{p-1} i.e. for all $\mathbf{u} \in S^{p-1}$, there exists at least one $\mathbf{v} \in \mathcal{C}^{p-1}$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \epsilon$. Standard constructions [17, see Lemma 5.2] guarantee such a cover of size at most $\left(1 + \frac{2}{\epsilon}\right)^p \leq \left(\frac{3}{\epsilon}\right)^p$. Now for any $\mathbf{u} \in S^{p-1}$ and $\mathbf{v} \in \mathcal{C}^{p-1}$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \epsilon$, we have

$$\left|\mathbf{u}^\top A\mathbf{u} - \mathbf{v}^\top A\mathbf{v}\right| \leq \left|\mathbf{u}^\top A(\mathbf{u} - \mathbf{v})\right| + \left|\mathbf{v}^\top A(\mathbf{u} - \mathbf{v})\right| \leq 2\epsilon \left\|A\right\|_2,$$

which gives us

$$\left\|XX^\top - nI\right\|_2 \leq (1 - 2\epsilon)^{-1} \cdot \sup_{\mathbf{v} \in \mathcal{C}^{p-1}(\epsilon)} \left|\left\|X^\top \mathbf{v}\right\|_2^2 - n\right|.$$

Now for a fixed $\mathbf{v} \in S^{n-1}$, the random variable $\left\|X^\top \mathbf{v}\right\|_2^2$ is distributed as a $\chi^2(n)$ distribution with n degrees of freedom. Using Lemma 20, we get, for any $\mu < 1$,

$$\mathbb{P}\left[\left|\left\|X^\top \mathbf{v}\right\|_2^2 - n\right| \geq \mu n\right] \leq 2\exp\left(-\min\left\{\frac{\mu^2 n^2}{24ne^2}, \frac{\mu n}{4\sqrt{3}e}\right\}\right) \leq 2\exp\left(-\frac{\mu^2 n}{24e^2}\right).$$

Setting $\mu^2 = c \cdot \frac{p}{n} + c' \cdot \frac{\log\frac{2}{\delta}}{n}$, where $c = 24e^2 \log\frac{3}{\epsilon}$ and $c' = 24e^2$, and taking a union bound over all $\mathcal{C}^{p-1}(\epsilon)$, we get

$$\mathbb{P}\left[\sup_{\mathbf{v} \in \mathcal{C}^{p-1}(\epsilon)} \left|\left\|X^\top \mathbf{v}\right\|_2^2 - n\right| \geq \sqrt{cnp + c'n\log\frac{2}{\delta}}\right] \leq 2\left(\frac{3}{\epsilon}\right)^p \exp\left(-\frac{\mu^2 n}{24e^2}\right) \leq \delta.$$

This implies that with probability at least $1 - \delta$,

$$\left\|XX^\top - nI\right\|_2 \leq (1 - 2\epsilon)^{-1}\sqrt{cnp + c'n\log\frac{2}{\delta}},$$

which gives us the claimed bounds on the singular values of $XX^\top$. □

Theorem 15. *Let $X \in \mathbb{R}^{p \times n}$ be a matrix whose columns are sampled i.i.d from a standard Gaussian distribution i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I)$. Then for any $\gamma > 0$, with probability at least $1 - \delta$, the matrix X satisfies the SSC and SSS properties with constants*

$$\Lambda_\gamma^{\textit{Gauss}} \leq \gamma n\left(1 + 3e\sqrt{6\log\frac{e}{\gamma}}\right) + \mathcal{O}\left(\sqrt{np + n\log\frac{1}{\delta}}\right)$$

$$\lambda_\gamma^{\textit{Gauss}} \geq n - (1-\gamma)n\left(1 + 3e\sqrt{6\log\frac{e}{1-\gamma}}\right) - \Omega\left(\sqrt{np + n\log\frac{1}{\delta}}\right).$$

Proof. For any fixed $S \in \mathcal{S}_\gamma$, Lemma 14 guarantees the following bound

$$s_{\max}(X_S X_S^\top) \leq \gamma n + (1 - 2\epsilon)^{-1}\sqrt{c\gamma np + c'\gamma n\log\frac{2}{\delta}}.$$

Taking a union bound over $\mathcal{S}_\gamma$ and noting that $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ for all $1 \leq k \leq n$, gives us

$$\begin{aligned}\Lambda_\gamma &\leq \gamma n + (1 - 2\epsilon)^{-1}\sqrt{c\gamma np + c'\gamma^2 n^2\log\frac{e}{\gamma} + c'\gamma n\log\frac{2}{\delta}} \\ &\leq \gamma n\left(1 + (1 - 2\epsilon)^{-1}\sqrt{c'\log\frac{e}{\gamma}}\right) + (1 - 2\epsilon)^{-1}\sqrt{c\gamma np + c'\gamma n\log\frac{2}{\delta}},\end{aligned}$$

which finishes the first bound after setting $\epsilon = 1/6$. For the second bound, we use the equality

$$X_S X_S^\top = XX^\top - X_{\bar{S}} X_{\bar{S}}^\top,$$

which provides the following bound for λ_γ

$$\lambda_\gamma \geq s_{\min}(XX^\top) - \sup_{T \in \mathcal{S}_{1-\gamma}} X_T X_T^\top = s_{\min}(XX^\top) - \Lambda_{1-\gamma}.$$

Using Lemma 14 to bound the first quantity and the first part of this theorem to bound the second quantity gives us, with probability at least $1 - \delta$,

$$\lambda_\gamma \geq n - \gamma' n \left(1 + (1-2\epsilon)^{-1} \sqrt{c' \log \frac{e}{\gamma'}}\right) - (1-2\epsilon)^{-1} \left(1 + \sqrt{\gamma'}\right) \sqrt{cnp + c' n \log \frac{2}{\delta}},$$

where $\gamma' = 1 - \gamma$. This proves the second bound after setting $\epsilon = 1/6$. $\square$

We now extend our analysis to the class of isotropic subGaussian distributions. We note that this analysis is without loss of generality since for non-isotropic sub-Gaussian distributions, we can simply use the fact that Theorem 3 can admit whitened data for calculation of the SSC and SSS constants as we did for the case of non-isotropic Gaussian distributions.

Lemma 16. *Let $X \in \mathbb{R}^{p \times n}$ be a matrix with columns sampled from some sub-Gaussian distribution with sub-Gaussian norm K and covariance Σ. Then, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following statements holds true:*

$$s_{\max}(XX^\top) \leq \lambda_{\max}(\Sigma) \cdot n + C_K \cdot \sqrt{pn} + t\sqrt{n}$$
$$s_{\min}(XX^\top) \geq \lambda_{\min}(\Sigma) \cdot n - C_K \cdot \sqrt{pn} - t\sqrt{n},$$

where $t = \sqrt{\frac{1}{c_K} \log \frac{2}{\delta}}$, and c_K, C_K are absolute constants that depend only on the sub-Gaussian norm K of the distribution.

Proof. Since the singular values of a matrix are unchanged upon transposition, we shall prove the above statements for $X^\top$. The benefit of this is that we get to work with a matrix with independent rows, so that standard results can be applied. The proof technique used in [17, Theorem 5.39] (see also Remark 5.40 (1) therein) can be used to establish the following result: with probability at least $1 - \delta$, with t set as mentioned in the theorem statement, we have

$$\left\| \frac{1}{n} XX^\top - \Sigma \right\| \leq C_K \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}$$

This implies that for any $\mathbf{v} \in S^{p-1}$, we have

$$\left| \frac{1}{n} \left\| X^\top \mathbf{v} \right\|_2^2 - \mathbf{v}^\top \Sigma \mathbf{v} \right| = \left| \frac{1}{n} \mathbf{v}^\top XX^\top \mathbf{v} - \mathbf{v}^\top \Sigma \mathbf{v} \right| \leq \left| \frac{1}{n} XX^\top \mathbf{v} - \Sigma \mathbf{v} \right| \leq C_K \sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}.$$

The results then follow from elementary manipulations and the fact that the singular values and eigenvalues of real symmetric matrices coincide. $\square$

Theorem 17. *Let $X \in \mathbb{R}^{p \times n}$ be a matrix with columns sampled from some sub-Gaussian distribution with sub-Gaussian norm K and covariance Σ. Let c_K, C_K and t be fixed to values as required in Lemma 16. Note that c_K and C_K are absolute constants depend only on the sub-Gaussian norm K of the distribution. Let $\gamma \in (0, 1]$ be some fixed constant. Then, with we have the following:*

$$\Lambda_\gamma^{\textit{subGauss}(K,\Sigma)} \leq \left(\lambda_{\max}(\Sigma) \cdot \gamma + \sqrt{\frac{\gamma}{c_K} \log \frac{e}{\gamma}}\right) \cdot n + C_K \cdot \sqrt{\gamma pn} + t\sqrt{n}.$$

Furthermore, fix any $\epsilon \in (0, 1)$ and let γ be a value in $(0, 1)$ satisfying the following

$$\gamma > 1 - \min\left\{ \frac{\epsilon \cdot \lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}, \exp\left(1 + W_{-1}\left(-\frac{c_K \epsilon^2 \cdot \lambda_{\min}^2(\Sigma)}{e}\right)\right)\right\},$$

where $W_{-1}(\cdot)$ is the lower branch of the real valued restriction of the Lambert W function. Then we have, with the same confidence,

$$\lambda_\gamma^{\textit{subGauss}(K,\Sigma)} \geq (1 - 2\epsilon) \cdot \lambda_{\min}(\Sigma) \cdot n - C_K \left(1 + \sqrt{1-\gamma}\right) \sqrt{pn} - 2t\sqrt{n}$$

Proof. The first result follows from an application of Lemma 16, a union bound over sets in $\mathcal{S}_\gamma$, as well as the bound $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ for all $1 \leq k \leq n$ which puts a bound on the number of sparse sets as $\log |\mathcal{S}_\gamma| \leq \gamma \cdot n \log \frac{e}{\gamma}$.

For the second result, we observe that $X_S X_S^\top = XX^\top - X_{\bar{S}} X_{\bar{S}}^\top$, so that $s_{\min}(X_S X_S^\top) \geq s_{\min}(XX^\top) - s_{\max}(X_{\bar{S}} X_{\bar{S}}^\top)$. This gives us

$$\inf_{S \in \mathcal{S}_\gamma} s_{\min}(X_S X_S^\top) \geq s_{\min}(XX^\top) - \sup_{S \in \mathcal{S}_{1-\gamma}} s_{\max}(X_S X_S^\top).$$

Using Lemma 16 and the first part of this result gives us

$$\begin{aligned}
\inf_{S \in \mathcal{S}_\gamma} s_{\min}(X_S X_S^\top) &\geq \lambda_{\min}(\Sigma) \cdot n - C_K \cdot \sqrt{pn} - t\sqrt{n} \\
&\quad - \left(\lambda_{\max}(\Sigma)(1-\gamma) + \sqrt{\frac{1-\gamma}{c_K} \log \frac{e}{1-\gamma}}\right) n - C_K \sqrt{(1-\gamma)pn} - t\sqrt{n} \\
&= \left(\lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma)(1-\gamma) - \sqrt{\frac{1-\gamma}{c_K} \log \frac{e}{1-\gamma}}\right) n \\
&\quad - C_K \left(1 + \sqrt{1-\gamma}\right) \sqrt{pn} - 2t\sqrt{n} \\
&\geq (1 - 2\epsilon) \cdot \lambda_{\min}(\Sigma) \cdot n - C_K \left(1 + \sqrt{1-\gamma}\right) \sqrt{pn} - 2t\sqrt{n},
\end{aligned}$$

where the last step follows from the assumptions on γ and by noticing that it suffices to show the following two inequalities to establish the last step

1. $\lambda_{\max}(\Sigma)(1-\gamma) \leq \epsilon \cdot \lambda_{\min}(\Sigma)$
2. $(1-\gamma) \log \frac{e}{1-\gamma} \leq c_K \epsilon^2 \cdot \lambda_{\min}^2(\Sigma)$

The first part gives us the condition $\gamma > 1 - \frac{\epsilon \cdot \lambda_{\min}(\Sigma)}{\lambda_{\max}(\Sigma)}$ in a straightforward manner. For the second part, denote $v = c_K \epsilon^2 \cdot \lambda_{\min}^2(\Sigma)$. Note that for $v \geq 1$, *all* values of $\gamma \in (0, 1]$ satisfy the inequality.

Otherwise we require the use of the Lambert W function (also known as the product logarithm function). This function ensures that its value $W(z)$ for any $z > -1/e$ satisfies $z = W(z)e^{W(z)}$. In our case, making a change of variable $(1-\gamma) = e^\eta$ gives us the inequality $(\eta - 1)e^{\eta - 1} \geq -v/e$. Note that since $v \leq 1$ in this case, $-v/e \in (-1/e, 0)$ i.e. a valid value for the Lambert W function. However, $(-1/e, 0)$ is also the region in which the Lambert W function is multi-valued. Taking the worse bound for γ by choosing the lower branch $W_{-1}(\cdot)$ gives us the second condition $\gamma \geq 1 - \exp\left(1 + W_{-1}\left(-\frac{c_K \epsilon^2 \cdot \lambda_{\min}^2(\Sigma)}{e}\right)\right)$. □

It is important to note that for any $-1/e \leq z < 0$, we have $\exp\left(1 + W_{-1}(z)\right) > 0$ which means that the bounds imposed on γ by Theorem 17 always allow a non-zero fraction of the data points to be corrupted in an adversarial manner. However, the exact value of that fraction depends, in a complicated manner, on the sub-Gaussian norm of the underlying distribution, as well as the condition number and the smallest eigenvalue of the second moment of the underlying distribution.

We also note that due to the generic nature of the previous analysis, which can handle the entire class of sub-Gaussian distributions, the bounds are not as explicitly stated in terms of universal constants as they are for the standard Gaussian design setting (Theorem 15).

We now establish that for a wide family of random designs, the SRSC and SRSS properties are satisfied with high probability as well. For sake of simplicity, we will present our analysis for the standard Gaussian design. However, the results would readily extend to general Gaussian and sub-Gaussian designs using techniques similar to Theorem 17.

Theorem 18. *Let $X \in \mathbb{R}^{p \times n}$ be a matrix whose columns are sampled i.i.d from a standard Gaussian distribution i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I)$. Then for any $\gamma > 0$ and $s \leq p$, with probability at least $1 - \delta$, the* *matrix X satisfies the SRSC and SRSS properties with constants*

$$L^{\textit{Gauss}}_{(\gamma,s)} \leq \gamma n\left(1 + 3e\sqrt{6\log\frac{e}{\gamma}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{ns + n\log\frac{1}{\delta}}\right)$$

$$\alpha^{\textit{Gauss}}_{(\gamma,s)} \geq n - (1-\gamma)n\left(1 + 3e\sqrt{6\log\frac{e}{1-\gamma}}\right) - \tilde{\Omega}\left(\sqrt{ns + n\log\frac{1}{\delta}}\right).$$

Proof. The proof of this theorem proceeds similarly to that of Theorem 15. Hence, we simply point out the main differences. First, we shall establish, that for any $\epsilon > 0$, with probability at least $1-\delta$, X satisfies the RSC and RSS properties at level s with the following constants

$$L_s \leq n + (1-2\epsilon)^{-1}\sqrt{bns + b'n\log\frac{2}{\delta}}$$

$$\alpha_s \geq n - (1-2\epsilon)^{-1}\sqrt{bns + b'n\log\frac{2}{\delta}},$$

where $b = 24e^2\log\frac{3ep}{\epsilon s}$ and $b' = 24e^2$. To do so we notice that the only change needed to be made would be in the application of the covering number argument. Instead of applying the union bound over an ϵ-cover $\mathcal{C}^{p-1}$ of S^{p-1}, we would only have to consider an ϵ-cover $\mathcal{C}^{p-1}_s$ of the set S^{p-1}_s of all s-sparse unit vectors in p-dimensions. A straightforward calculation shows us that

$$\left|\mathcal{C}^{p-1}_s\right| \leq \binom{p}{s}\left(1+\frac{2}{\epsilon}\right)^s \leq \left(\frac{3ep}{\epsilon s}\right)^s.$$

Thus, setting $\mu^2 = b\cdot\frac{s}{n} + b'\cdot\frac{\log\frac{2}{\delta}}{n}$, where $b = 24e^2\log\frac{3ep}{\epsilon s}$ and $b' = 24e^2$, we get

$$\mathbb{P}\left[\sup_{\mathbf{v}\in\mathcal{C}^{p-1}_s}\left|\|X\mathbf{v}\|_2^2 - n\right| \geq \sqrt{bns + b'n\log\frac{2}{\delta}}\right] \leq \delta,$$

which establishes the required RSC and RSS constants for X. Now, moving on to the SRSS constant, it follows simply by applying a union bound over all sets in $\mathcal{S}_\gamma$ much like in Theorem 15. One can then proceed to bound the SRSC constant in a similar manner.

We note that the nature of the SRSC and SRSS bounds indicate that our TORRENT-FC algorithm in the high dimensional sparse recovery setting has noise tolerance properties, characterized by the largest corruption index α that can be tolerated, identical to its low dimnensional counterpart - something that Theorem 9 states explicitly. □

H Supplementary Results

Claim 19. *Given any vector $\mathbf{v} \in \mathbb{R}^n$, let $\sigma \in S_n$ be defined as the permutation that orders elements of $\mathbf{v}$ in descending order of their magnitudes i.e. $\left|v_{\sigma(1)}\right| \geq \left|v_{\sigma(2)}\right| \geq \ldots \geq \left|v_{\sigma(n)}\right|$. For any $0 < p \leq q \leq 1$, let $S_1 \in \mathcal{S}_q$ be an arbitrary set of size $q\cdot n$ and $S_2 = \{\sigma(i) : n - p\cdot n + 1 \leq i \leq n\}$. Then we have $\|\mathbf{v}_{S_2}\|_2^2 \leq \frac{p}{q}\|\mathbf{v}_{S_1}\|_2^2 \leq \|\mathbf{v}_{S_1}\|_2^2$.*

Proof. Let $S_3 = \{\sigma(i) : n - q\cdot n + 1 \leq i \leq n\}$ and $S_4 = \{\sigma(i) : n - q\cdot n + 1 \leq i \leq n - p\cdot n\}$. Clearly, we have $\|\mathbf{v}_{S_3}\|_2^2 \leq \|\mathbf{v}_{S_1}\|_2^2$ since S_3 contains the smallest $q\cdot n$ elements (by magnitude). Now we have $\|\mathbf{v}_{S_3}\|_2^2 = \|\mathbf{v}_{S_2}\|_2^2 + \|\mathbf{v}_{S_4}\|_2^2$. Moreover, since each element of S_4 is larger in magnitude than every element of S_2, we have

$$\frac{1}{|S_4|}\|\mathbf{v}_{S_4}\|_2^2 \geq \frac{1}{|S_2|}\|\mathbf{v}_{S_2}\|_2^2.$$

This gives us

$$\|\mathbf{v}_{S_2}\|_2^2 = \|\mathbf{v}_{S_3}\|_2^2 - \|\mathbf{v}_{S_4}\|_2^2 \leq \|\mathbf{v}_{S_3}\|_2^2 - \frac{|S_4|}{|S_2|}\|\mathbf{v}_{S_2}\|_2^2,$$

which upon simple manipulations, gives us the claimed result. □

Lemma 20. *Let Z be distributed according to the chi-squared distribution with k degrees of freedom i.e. $Z \sim \chi^2(k)$. Then for all $t \geq 0$,*

$$\mathbb{P}\left[|Z - k| \geq t\right] \leq 2\exp\left(-\min\left\{\frac{t^2}{24ke^2}, \frac{t}{4\sqrt{3}e}\right\}\right)$$

Proof. This lemma requires a proof structure that traces several basic results in concentration inequalities for sub-exponential variables [17, Lemma 5.5, 5.15, Proposition 5.17]. The purpose of performing this exercise is to explicate the constants involved so that a crisp bound can be provided on the corruption index that our algorithm can tolerate in the standard Gaussian design case.

We first begin by establishing the sub-exponential norm of a chi-squared random variable with a single degree of freedom. Let $X \sim \chi^2(1)$. Then using standard results on the moments of the standard normal distribution gives us, for all $p \geq 2$,

$$(\mathbb{E}|X|^p)^{1/p} = ((2p-1)!!)^{1/p} = \left(\frac{(2p)!}{2^p p!}\right)^{1/p} \leq \frac{\sqrt{3}}{2}p$$

Thus, the sub-exponential norm of X is upper bounded by $\sqrt{3}/2$. By applying the triangle inequality, we obtain, as a corollary, an upper bound on the sub-exponential norm of the centered random variable $Y = X - 1$ as $\|Y\|_{\psi_1} \leq 2\|X\|_{\psi_1} \leq \sqrt{3}$.

Now we bound the moment generating function of the random variable Y. Noting that $\mathbb{E}Y = 0$, we have, for any $|\lambda| \leq \frac{1}{2\sqrt{3}e}$,

$$\mathbb{E}\exp(\lambda Y) = 1 + \sum_{q=2}^{\infty}\frac{\mathbb{E}(\lambda Y)^q}{q!} \leq 1 + \sum_{q=2}^{\infty}\frac{(\sqrt{3}|\lambda|q)^q}{q!} \leq 1 + \sum_{q=2}^{\infty}(\sqrt{3}e|\lambda|)^q \leq 1 + 6e^2\lambda^2 \leq \exp(6e^2\lambda^2).$$

Note that the second step uses the sub-exponentially of Y, the third step uses the fact that $q! \geq (q/e)^q$, and the fourth step uses the bound on $|\lambda|$. Now let $X_1, X_2, \ldots X_k$ be k independent random variables distributed as $\chi^2(1)$. Then we have $Z \sim \sum_{i=1}^k X_i$. Using the exponential Markov's inequality, and the independence of the random variables X_i gives us

$$\mathbb{P}\left[Z - k \geq t\right] = \mathbb{P}\left[e^{\lambda(Z-k)} \geq e^{\lambda t}\right] \leq e^{-\lambda t}\mathbb{E}e^{\lambda(Z-k)} = e^{-\lambda t}\prod_{i=1}^k \mathbb{E}\exp(\lambda(X_i - 1)).$$

For any $|\lambda| \leq \frac{1}{2\sqrt{3}e}$, the above bounds on the moment generating function give us

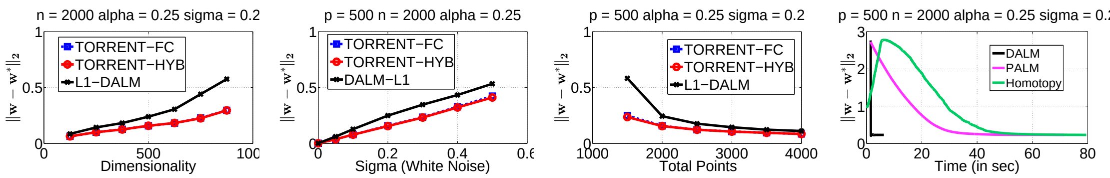
$$\mathbb{P}\left[Z - k \geq t\right] \leq e^{-\lambda t}\prod_{i=1}^k \exp(6e^2\lambda^2) = \exp(-\lambda t + 6ke^2\lambda^2).$$

Choosing $\lambda = \min\left\{\frac{1}{2\sqrt{3}e}, \frac{t}{12ke^2}\right\}$, we get

$$\mathbb{P}\left[Z - k \geq t\right] \leq \exp\left(-\min\left\{\frac{t^2}{24ke^2}, \frac{t}{4\sqrt{3}e}\right\}\right).$$

Repeating this argument gives us the same bound for $\mathbb{P}\left[k - Z \geq t\right]$. This completes the proof. □

I Supplementary Experimental Results



(a) (b) (c) (d)

Figure 3: (a), (b), (c) Variation of recovery error with varying p, σ and n. TORRENT was found to outperform DALM-L_1 in all these settings. (d) Recovery error as a function of runtime for various state-of-the-art L_1 solvers as indicated in [14].