Globally-convergent Iteratively Reweighted Least Squares for Robust Regression Problems

Bhaskar Mukhoty[†]   Govind Gopakumar[†*]   Prateek Jain[‡]

Purushottam Kar[†]

[†]IIT Kanpur

[‡]Microsoft Research India

{bhaskarm,govindg,purushot}@cse.iitk.ac.in, prajain@microsoft.com

June 26, 2020

Abstract

We provide the first global model recovery results for the IRLS (iteratively reweighted least squares) heuristic for robust regression problems. IRLS is known to offer excellent performance, despite bad initializations and data corruption, for several parameter estimation problems. Existing analyses of IRLS frequently require careful initialization, thus offering only local convergence guarantees. We remedy this by proposing augmentations to the basic IRLS routine that not only offer guaranteed global recovery, but in practice also outperform state-of-the-art algorithms for robust regression. Our routines are more immune to hyperparameter misspecification in basic regression tasks, as well as applied tasks such as linear-armed bandit problems. Our theoretical analyses rely on a novel extension of the notions of strong convexity and smoothness to *weighted strong convexity and smoothness*, and establishing that sub-Gaussian designs offer bounded *weighted condition numbers*. These notions may be useful in analyzing other algorithms as well.

1   Introduction

Suppose there exists an unknown gold model $\mathbf{w}^*$ and we are given n data points $(\mathbf{x}_i, y_i)_{i=1}^n$ with d-dimensional covariates $\mathbf{x}_i \in \mathbb{R}^d$ and the real-valued responses y_i generated as $y_i = \mathbf{x}_i^\top \mathbf{w}^*$. However, for an unknown set of $k < n$ data points $i_1, \ldots i_k$, the responses get corrupted i.e. we instead receive $y_{i_j} = \mathbf{x}_{i_j}^\top \mathbf{w}^* + b_{i_j}$ where $b_{i_j} \in \mathbb{R}$ is the corruption. Given the complete set of clean and corrupted data points $(\mathbf{x}_i, y_i)_{i=1}^n$, can we recover the gold model $\mathbf{w}^*$?

This is the classical robust regression problem that has become increasingly relevant to machine learning and statistical estimation techniques which frequently encounter situations where data is not trustworthy. Works exist in settings where test data is corrupted in order to fool a model that was learnt on clean data [17], as well as the more challenging setting, on which we focus, where the training data presented to the algorithm is itself corrupted [9, 11, 16].

We will seek to offer reliable model recovery despite the presence of (possibly maliciously) corrupted data in the training set. Settings which present corrupted data to learning algorithms include relatively innocuous instances of erasures and missing data, improperly or mistakenly attributed data, transient or temporary changes in user-behavior patterns, as well as deliberate and malicious attempts to derail recommendation systems and other decision-making systems using malware, click-bots and other fraudulent techniques.

Despite being a well established field, given the early seminal contributions of Huber [18] and Tukey [27], robust statistics and algorithms have received renewed interest given the threat to modern machine learning

---

[*]Work done as a master's student at IIT Kanpur

1

Table 1: Algorithms for the Robust Regression problem (corrupted responses). [†]Please see §4 for details. Algorithms able to tolerate adaptive (as opposed to oblivious) adversaries are more resilient. A more robust algorithm can handle larger α. Sub-Gaussian covariates offer a much more flexible model than (isotropic) Gaussian covariates.

Paper	Adversary Model[†]	Breakdown point[†]	Covariate Model	Technique
Bhatia et. al. 2015 [6]	Adaptive	$\alpha \geq \Omega(1)$	sub-Gaussian	Hard Thresholding (fast)
Chen & Dalalyan 2010 [10]	Adaptive	$\alpha \geq \Omega(1)$	sub-Gaussian	SOCP (slow)
Wright & Ma 2010 [29]	Oblivious	$\alpha \to 1$	Isotropic Gaussian	L_1 regularization (slow)
This Paper	**Adaptive**	$\boldsymbol{\alpha \geq \Omega(1)}$	**sub-Gaussian**	**Reweighting (fast)**

techniques. Of the several techniques that have been proposed for robust learning problems, one heuristic, namely the iteratively reweighted least squares (IRLS), remains a practitioner's favorite owing to its ease of use and excellent performance. The IRLS technique has been effectively adapted to several problems, including sparse recovery, and robust regression. The work of [26] shows that certain biological dynamical systems can be modeled upon the IRLS principle as well.

1.1 Our Contributions

We offer several advances in the understanding and application of the IRLS method. In particular, we provide the first global model recovery guarantee for IRLS for robust regression - our contributions are distinguished in the context of existing analyses for IRLS in §2. We also propose algorithmic augmentations, in particular a fast gradient-based variant, to the basic IRLS heuristic which offer superior performance compared to existing state-of-the-art robust algorithms in terms of speed, as well as resilience to misspecified hyperparameters. We demonstrate this in the standard linear regression setting, as well as an applied setting, namely linear-armed bandits.

2 Related Work

Two lines of work directly relate to our contributions: 1) robust algorithms for regression and other learning problems, and 2) works that analyze (variants of) the IRLS heuristic in various settings. We review both, as well as distinguish our contributions, below.

Robust Learning Algorithms: Work on robust statistics dates back several decades [18, 27] and is too vast to be reviewed in detail. Recent years have seen interest in scalable algorithms for classification [16], principal component analysis [9], and moment estimation [14]. Within the specific problem of robust regression, two broad lines of work exist:

Covariate (feature) corruption: Results in this setting usually either give only weak guarantees, or else severely constrain data. e.g., [11, 24] allow only a $\mathcal{O}\left(1/\sqrt{d}\right)$ fraction of data to be corrupted, d being the ambient dimensionality, whereas [15, 21] only admit covariates drawn from a Gaussian distribution.

Response (label) corruption: Variants within this setting arise based on the power of the adversary introducing the corruptions, the fraction of data points that can be corrupted, restrictions on the choice of covariates, and scalability of the algorithms. Table 1 summarizes these traits for a selection of algorithms. We refer the reader to [6, 15] for other references.

IRLS Variants and Analyses: The IRLS heuristic has been successfully applied to several problems including sparse recovery [4, 13], facility location problems [8] (via the Weiszfeld procedure), and optimizing various robust cost functions, such as the L_q and Huber loss functions [2, 7, 12, 25].

Some of these works are not directly relevant to robust regression as they either operate with uncorrupted data [8], or else assume that the noise is Gaussian [4, 13]. Convergence guarantees for IRLS are common in these benign settings. To handle adversarial corruptions, it is common to use IRLS to optimize a *robust cost function* F such as L_q or Huber loss, in the anticipation that the model so obtained, say $\hat{\mathbf{w}} = \arg\min F(\mathbf{w}; \{(\mathbf{x}_i, y_i)\})$, will ensure $\hat{\mathbf{w}} \approx \mathbf{w}^*$.

However, none of these works actually ensure such a result i.e. $\hat{\mathbf{w}} \approx \mathbf{w}^*$. Some works [7, 12, 25] operate with cost functions that are convex (e.g. L_q for $q \in [1, 2]$) and simply show that IRLS approaches small cost function values. Other approaches [2] do work with non-convex cost functions, but then offer only monotonicity guarantees and no global convergence guarantees.

We bridge this gap by presenting a much stronger analysis of IRLS that guarantees *global recovery* of the gold model $\mathbf{w}^*$ under mild conditions. Key to our proof technique is a novel concept that extends the basic notions of strong convexity and strong smoothness to *weighted* versions of the same, as well as a guarantee that Gaussian and sub-Gaussian designs have bounded *weighted condition numbers*. These results may be of independent interest in analyzing other algorithms.

3 Notation

Bold lower-case Latin letters $\mathbf{x}, \mathbf{y}$ denote vectors. $\mathbf{x}_i$ denotes the ith coordinate of the vector $\mathbf{x}$. Upper case Latin letters A, X denote matrices. For a vector $\mathbf{v} \in \mathbb{R}^n$ and set $S \subset [n]$, $\mathbf{v}_S$ denotes the vector with $(\mathbf{v}_S)_i = \mathbf{v}_i$ for $i \in S$ and $(\mathbf{v}_S)_j = 0$ for $j \notin S$. Similarly, for any matrix $A \in \mathbb{R}^{d \times n}$ and any set $S \subset [n]$, A_S denotes the matrix in which columns $i \in S$ in A_S are identical to those in A and columns $j \notin S$ are filled with zeros.

$\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ denote, respectively, the smallest and largest eigenvalues of a square symmetric matrix X. $\mathcal{B}_2(\mathbf{v}, r) := \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{v}\|_2 \leq r \right\}$ denotes the ball of radius r centered at $\mathbf{v}$. S^{d-1} denotes the surface of the unit sphere in d dimensions. We use the shorthand $\mathcal{B}_2(r) := \mathcal{B}_2(\mathbf{0}, r)$.

4 Problem Formulation

Given n data points $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, let $R_X := \max_{i \in [n]} \|\mathbf{x}_i\|_2$ be the maximum Euclidean length of any covariate, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ be the covariate matrix, and $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ the response vector. Assume that the covariates are generated as $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{D}$ from an unknown distribution $\mathcal{D}$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and sub-Gaussian norm [28] $\|\mathcal{D}\|_{\Psi_2} \leq R$. $\mathbf{w}^* \in \mathbb{R}^d$ will be the gold model with $R_W := \|\mathbf{w}^*\|_2$.

Noise Model: Given the data covariates and the gold model, the responses are generated as $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$ where $\mathbf{b} = [b_1, \ldots, b_n]$ is the vector of corruptions. We make the standard assumption that $\|\mathbf{b}\|_0 \leq \alpha \cdot n$. Let $B := \text{supp}(\mathbf{b})$ denote the "bad" points which suffer corruption i.e. $\mathbf{b}_j \neq 0$ for $j \in B$ (note that $|B| \leq \alpha \cdot n$) and $G = [n] \setminus B$ denote the "good" points where $\mathbf{b}_i = 0$ and thus $y_i = \mathbf{x}_i^\top \mathbf{w}^*$ for $i \in G$. To avoid clutter, we abuse notation to denote $G := |G|$ and $B := |B|$. The largest value of the corruption fraction α that an algorithm can tolerate is known as its *breakdown point*.

Adversary Model: We will work with a partially adaptive adversary which is compelled to choose locations of the corruptions $\text{supp}(\mathbf{b}) = B$ before any data covariates have been generated or $\mathbf{w}^*$ is revealed. However, the adversary may fill in the corruption values at those locations with knowledge of $\mathbf{w}^*$ and X. Our results can be extended to a *fully adaptive adversary* that choose $\text{supp}(\mathbf{b})$ after looking at $\mathbf{w}^*$ and X as well, but at a cost of a smaller breakdown point α.

Key to our analyses are the notions of *weighted strong convexity and smoothness* which we define below. These definitions reflect the fact that IRLS solves *weighted* regression problems iteratively.

Definition 1 (WSC/WSS). *We say that a covariate matrix $X \in \mathbb{R}^{d \times n}$ offers* weighted strong convexity *(WSC) at level λ_S (resp.* weighted strong smoothness *(WSS) at level Λ_S), with respect to a diagonal weight matrix $S = diag(\mathbf{s}) \in \mathbb{R}^{n \times n}$ where $\mathbf{s}_i \geq 0, i \in [n]$, if*

$$\lambda_S \leq \lambda_{\min}(XSX^\top) \leq \lambda_{\max}(XSX^\top) \leq \Lambda_S$$

5 Proposed Methods

IRLS solves the robust regression problem by repeatedly alternating between the following two steps

1. **Reweighing**: Given a model $\hat{\mathbf{w}}$, assign every data point a weight s_i inversely proportional to its residual w.r.t. $\hat{\mathbf{w}}$ i.e. set $\mathbf{s}_i = \frac{1}{\left|\mathbf{x}_i^\top \hat{\mathbf{w}} - y_i\right|}$.

2. **Weighted Least Squares**: Solve a weighted least squares problem $\min_{\mathbf{w}} \sum_{i=1}^n \mathbf{s}_i (y_i - \mathbf{x}_i^\top \mathbf{w})^2$ with above weights to obtain a new model $\mathbf{w}^+ = (XSX^\top)^{-1} XS\mathbf{y}$ where $S = \mathrm{diag}(\mathbf{s})$.

The intuition behind this procedure is that corrupted points are likely to suffer large residuals and hence get downweighted. Given that this procedure runs the risk of divide-by-zero errors and numerical precision issues, it is common to truncate weights by employing a *truncation parameter* M while assigning weights[1] to the points i.e. $\mathbf{s}_i = \min\left\{\frac{1}{\left|\mathbf{x}_i^\top \hat{\mathbf{w}} - y_i\right|}, M\right\}$. However, it is suboptimal to rely on any single truncation value M. To see why, take a hypothetical example where the adversary introduces corruptions using a *fake model* $\tilde{\mathbf{w}}$ as $b_i = \mathbf{x}_i^\top (\tilde{\mathbf{w}} - \mathbf{w}^*)$ (i.e. $y_i = \mathbf{x}_i^\top \tilde{\mathbf{w}}$) for all $i \in B$.

Situation 1: If we set M to a small value (aggressive truncation), then no data point can ever hope to get a large weight. However, convergence to $\mathbf{w}^*$ is assured only when points in G receive really large weights in comparison to points in B. Setting a small value of M thus prevents IRLS from recovering $\mathbf{w}^*$ accurately.

Situation 2: If we always use a large value of M (lax truncation) and are unlucky enough to initialize IRLS close to $\tilde{\mathbf{w}}$, then points in the set B will initially have very small residuals, hence receive large weights (which the large value of M will allow) whereas points in the set G will receive comparatively smaller weights. This will cause IRLS to gravitate towards $\tilde{\mathbf{w}}$. This example precludes any hope of a global convergence guarantee and forces us to do careful initialization.

The above limitations of IRLS are well corroborated by experiments (see §8). To remedy this, we propose the STIR algorithm in Algorithm 1. STIR executes IRLS, but in *stages*, with initial stages employing aggressive truncation with a small value of M and later stages successively relaxing the truncation.

The advantage of the above augmentation is that even if we have an unfortunate initialization, e.g. we start at $\tilde{\mathbf{w}}$ itself, the (initially) aggressive truncation will prevent bad points from getting large weights whereas good points, being in majority, even though receiving relatively smaller weights, will still prevent STIR from latching onto $\tilde{\mathbf{w}}$ and hopefully attract the procedure towards the gold model $\mathbf{w}^*$. Subsequent stages, where truncation is relaxed, will allow good points to be given large weights, thus differentiating them from bad points. This would force STIR towards $\mathbf{w}^*$.

Algorithm 2 presents STIR-GD, a gradient version of STIR, that replaces weighted least squares by a much cheaper gradient step. This benefits large datasets, where solving weighted least squares repeatedly may be prohibitive. We note that although *stagewise* IRLS procedures have been proposed in literature [7], previous works neither give model recovery guarantees, nor offer scalable gradient versions of IRLS.

6 IRLS is Majorization-minimization on a Scaled Huber Loss

Before presenting a convergence analysis for STIR, we point out a curious link between IRLS, STIR and the Huber loss function. We note that our observation may be folklore. The Huber loss is widely used in robust regression applications [2, 7, 12, 25], particularly those used in situations with heavy tailed noise.

$$h_\epsilon(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le \epsilon \\ \epsilon |x| - \frac{1}{2}\epsilon^2 & |x| > \epsilon \end{cases}$$

The function smoothly transitions from quadratic behavior close to the origin, to linear far from the origin. Now consider the following loss function

$$f_\epsilon(x) = \begin{cases} \frac{1}{2}\left(\frac{x^2}{\epsilon} + \epsilon\right) & |x| \le \epsilon \\ |x| & |x| > \epsilon \end{cases}$$

---

[1]Literature often cites a *regularization* procedure that sets $\mathbf{s}_i = \frac{1}{\max\{|\mathbf{x}_i^\top \hat{\mathbf{w}} - y_i|, \delta\}}$ given a parameter δ. Setting $\delta = \frac{1}{M}$ shows truncation to be equivalent to regularization.

Algorithm 1 STIR- Stagewise-Truncated IRLS

Input: Data $X, \mathbf{y}$, initial truncation M_1, increment $\eta > 1$
Output: A model $\mathbf{w}$
  1: $\mathbf{w}^1 \leftarrow \mathbf{0}$
  2: **for** $T = 1, 2, \ldots, K - 1$ **do**
  3:    $\mathbf{w}^{T,1} \leftarrow \mathbf{w}^T$
  4:    $t \leftarrow 1$
  5:    **while** $\left\| \mathbf{w}^{T,t+1} - \mathbf{w}^{T,t} \right\|_2 > \frac{2}{\eta M_T}$ **do**
  6:        $\mathbf{r}^t \leftarrow X^\top \mathbf{w}^{t,1} - \mathbf{y}$
  7:        $S^t \leftarrow \mathrm{diag}(\mathbf{s}^t), \qquad \mathbf{s}_i^t \leftarrow \min \left\{ \frac{1}{|\mathbf{r}_i^t|}, M_T \right\}$
  8:        $\mathbf{w}^{T,t+1} \leftarrow (X S^t X^\top)^{-1} X S^t \mathbf{y}$
  9:        $t \leftarrow t + 1$
10:    **end while**
11:    $\mathbf{w}^{T+1} \leftarrow \mathbf{w}^{T,t+1}$
12:    $M_{T+1} \leftarrow \eta \cdot M_T$
13: **end for**
14: **return** $\mathbf{w}^K$

Algorithm 2 STIR-GD: STIR-Gradient Descent

Input: Data $X, \mathbf{y}$, initial truncation M_1, increment $\eta > 1$, step length C
Output: A model $\mathbf{w}$
  8:    $\mathbf{w}^{T,t+1} \leftarrow \mathbf{w}^{T,t} - \frac{2C}{M_T n} \cdot X S^t \mathbf{r}^t$

                                          `//Rest of steps 1-14 remain same as in STIR`

It is easily seen that $f_\epsilon(x) = \frac{h_\epsilon(x)}{\epsilon} + \frac{\epsilon}{2}$ and thus, $f_\epsilon()$ is simply a scaled (and translated) version of the Huber loss function, as well as that $|x| \leq f_\epsilon(x) \leq |x| + \frac{\epsilon}{2}$. Now, for any $a \in \mathbb{R}, \epsilon > 0$, consider the function

$$g_\epsilon(x; a) := \frac{1}{2} \left(\frac{x^2}{\max\{|a|, \epsilon\}} + \max\{|a|, \epsilon\} \right)$$

Given a model $\mathbf{w}^0$ and data $(\mathbf{x}_i, y_i)_{i=1}^n$, denote

$$\ell_\epsilon(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n f_\epsilon \left(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i \right)$$

$$\wp_\epsilon(\mathbf{w}; \mathbf{w}^0) := \sum_{i=1}^n g_\epsilon \left(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i; \langle \mathbf{w}^0, \mathbf{x}_i \rangle - y_i \right)$$

The following observations are key (see Appendix A).

1. $\wp_\epsilon(\cdot; \mathbf{w}^0)$ is a majorizer for $\ell_\epsilon(\cdot)$ at $\mathbf{w}^0, \forall \epsilon > 0$ i.e. $\wp_\epsilon(\mathbf{w}; \mathbf{w}^0) \geq \ell_\epsilon(\mathbf{w}), \forall \mathbf{w}$ but $\wp_\epsilon(\mathbf{w}^0; \mathbf{w}^0) = \ell_\epsilon(\mathbf{w}^0)$

2. If the current model is $\mathbf{w}^0$ then M-truncated IRLS minimizes $\wp_{\frac{1}{M}}(\mathbf{w}; \mathbf{w}^0)$ to obtain the next model.

3. $\nabla \wp_\epsilon(\mathbf{w}^0; \mathbf{w}^0) = \nabla \ell_\epsilon(\mathbf{w}^0)$.

Thus, IRLS can be seen as performing majorization-minimization [23] on the scaled Huber loss $\ell_\epsilon(\cdot)$. The reweighing step effectively constructs the majorizer function $\wp_\epsilon(\cdot, \mathbf{w}^0)$ over which the least squares step then performs minimization. Point 3 above shows that STIR-GD can be effectively seen as performing gradient descent with respect to $\ell_\epsilon(\mathbf{w}^0)$.

This also allows us to interpret the stages of STIR as using scaled Huber losses with successively smaller values of ϵ (point 2 above shows that STIR sets $\epsilon = \frac{1}{M}$). Note that in the limit $\epsilon \to 0$, $\ell_\epsilon(\cdot)$ approaches the absolute error function, and thus, in the limit $M \to \infty$, STIR ends up optimizing the absolute error function. STIR-GD can be seen as simply replacing the minimization steps with a gradient descent step.
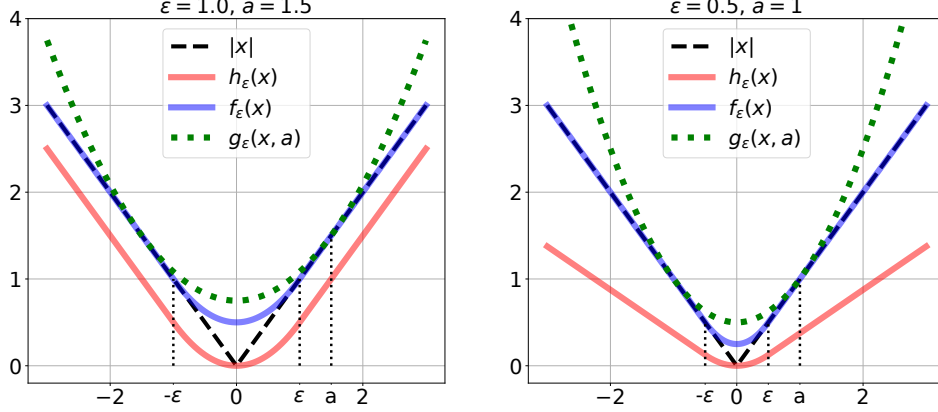
Figure 1: A depiction of Huber $h_\epsilon()$, scaled Huber $f_\epsilon()$ loss functions, and its majorizer $g_\epsilon()$ for various ϵ.

7 Convergence Analysis

In this section, we establish that both STIR and STIR-GD enjoy a linear rate of convergence, as well as a breakdown point $\alpha \geq \Omega(1)$. Theorem 1 summarizes the results. It is notable that STIR and STIR-GD offer a breakdown point of greater than $\frac{1}{5.25}$ (for Gaussian covariates – see below for details), which is far superior to those offered by recent works such as [6, 5] which offer breakdown points of $\approx \frac{1}{60}$ and $\frac{1}{10000}$ respectively (again for Gaussian covariates).

Theorem 1. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \leq \frac{c}{2.88\eta + c}$, where $c > 0$ is a constant that depends only on $\mathcal{D}$, then for any $\epsilon > 0$, with probability at least $1 - \exp(-\tilde{\Omega}(n))$, after $K = \mathcal{O}\left(\log \frac{1}{M_1 \epsilon}\right)$ stages, we must have $\left\|\mathbf{w}^K - \mathbf{w}^*\right\|_2 \leq \epsilon$. Moreover, each stage consists of only $\mathcal{O}(1)$ iterations.*

Global Convergence Note that the above result allows initialization at any location $\mathbf{w}^0$, so long as the accompanying value M_1 is small enough i.e. $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$ which can be ensured using a simple binary search (see §8 for details on parameter setting). In particular, if an estimated upper-bound $\|\mathbf{w}^*\|_2 \leq W$ is available, then we can set $\mathbf{w}^0 = \mathbf{0}$ and set $M_1 = \frac{1}{W}$.

Given this parameter convergence result, we can also establish that STIR and STIR-GD offer linear convergence guarantees with respect to the Huber and absolute loss functions as well. We refer the reader to Appendix C.2 for details.

Breakdown Point Both STIR and STIR-GD enjoy a breakdown point of $\alpha \leq \frac{c}{2.88\eta + c}$ where η is chosen by us and c is a distribution dependent constant. Bounds on this constant are established for several interesting distributions in Appendix D.1. In particular, for the Gaussian distribution $\mathcal{N}(\mathbf{0}, I_d)$, we have $c \geq 0.68$ which, for values of $\eta \to 1$, endow STIR and STIR-GD with a breakdown point of greater than $\frac{1}{5.25}$.

7.1 Proof Outline - the Peeling Strategy

Given the stage-wise nature of our algorithms STIR and STIR-GD, we employ a *peeling*-based proof strategy that is a departure from the techniques used by previous results such as [6, 10, 29].

Our proof partitions the model space into *annular peels* centered at the gold model $\mathbf{w}^*$ (see Figure 2). The outermost peel has a radius of $\frac{1}{M_1}$, and successive inner peels have radii that are an η factor smaller i.e. the subsequent peels have radii $\frac{1}{\eta M_1}, \frac{1}{\eta^2 M_1}, \frac{1}{\eta^3 M_1}, \dots$. Note that by setting $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, STIR is guaranteed to reside inside the outermost peel in the beginning.
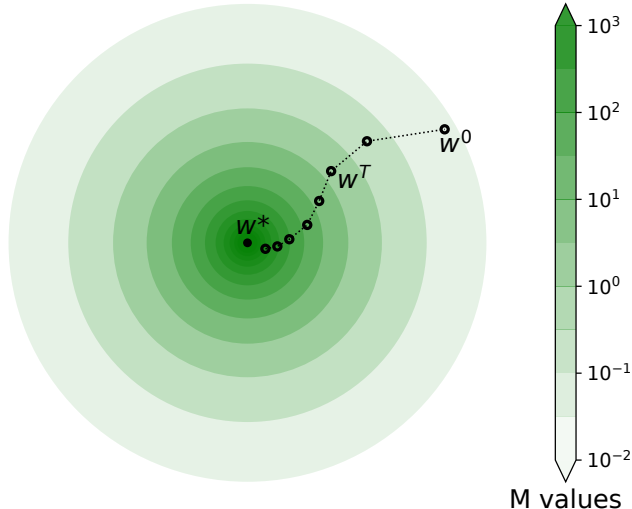
Figure 2: A depiction of the peeling process. The STIR procedure starts off far away from $\mathbf{w}^*$ and using a small value of M. In successive stages, it enters closer peels around $\mathbf{w}^*$ and also begins using larger values of M.

We then inductively show (see Lemmata 8 and 9) that once we are inside a certain peel, say $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{\eta^K M}$, and if the WSC/WSS properties hold with appropriate constants (see Appendix D), then if we execute $(\eta^K M)$-truncated IRLS for a constant number of iterations, we are guaranteed to obtain a model, say $\mathbf{w}^+$, that ensures $\|\mathbf{w}^+ - \mathbf{w}^*\|_2 \leq \frac{1}{\eta^{K+1} M}$.

This implies that we have entered the next inner peel. We can now set the truncation level to $\eta^{K+1} M$ and continue the process. Note that this is exactly the algorithmic step performed by STIR/STIR-GD (see Algorithm 1, line 12) to start a new stage. Due to lack of space, all complete proofs are given in the appendices.

7.2 Establishing WSC/WSS

A central result required for the peeling strategy to work, is ensuring that our covariates satisfy the WSC/WSS properties (that we introduced in §4) with respect to the weights assigned to data points by the STIR and STIR-GD algorithms. We show that for covariates drawn from sub-Gaussian distributions, this is indeed true (see Appendix D).

The use of such *design properties* is quite common in literature e.g., restricted strong convexity/smoothness (RSC/RSS) [13] in sparse recovery, and subset strong convexity/smoothness (SSC/SSS) [6] in robust regression. It is also common to use results on extremal singular values of random matrices [28], to show that sub-Gaussian covariates satisfy RSC/RSS [3] and SSC/SSS [6], with high probability.

However, doing so in our case is not as straightforward. The reason for this is that whereas the RSC/RSS and SSC/SSS properties are defined purely in terms of the data covariates, the WSC/WSS properties also incorporate data weights. Moreover, these weights are neither constant, nor independent of the data, but rather are assigned and repeatedly updated in a stage-wise manner by an algorithm such as IRLS or STIR.

Since our proofs will require the WSC/WSS properties to hold with respect to *all* weight assignments made during the entire execution of the algorithms, a direct application of classical techniques [28] fails. Such techniques could have succeeded only if the data weights were to be constant or else independent of the data.

To overcome this challenge, we establish WSC/WSS properties for sub-Gaussian covariates in a *peel-wise* manner using a careful uniform convergence bound. The number of peels is no more than $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ since each peel corresponds to a stage of the algorithm and $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ is the number of stages required to achieve

an ϵ-accurate solution (see Theorem 1), which then allows us to take a union bound over all peels.

Within each peel, a careful uniform convergence bound is employed over all models within that peel in order to establish WSC/WSS. Note that our results present a novel extension of the existing notions of SSC/SSS since we can recover SSC/SSS as a special case of WSC/WSS where the weights are simply zero or unity.

7.3 Corruptions and Dense Noise

So far we have looked at an idealized setting where the responses are either completely clean $y_i = \mathbf{x}_i^\top \mathbf{w}^*$ for $i \in G$ or else corrupted $y_j = \mathbf{x}_j^\top \mathbf{w}^* + \mathbf{b}_j$ for $j \in B$. We now look at a more realistic setting where even the "good" points experience sub-Gaussian noise. We will now assume that our data is generated as $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b} + \boldsymbol{\epsilon}$ where, as before $\|\mathbf{b}\|_0 \leq \alpha \cdot n$, but we additionally have $\boldsymbol{\epsilon} \sim \mathcal{D}_\varepsilon$ where $\mathcal{D}_\varepsilon$ is a σ-sub-Gaussian distribution with zero mean and real support [2].

We will denote $B := \text{supp}(\mathbf{b})$ and $G := [n] \setminus B$, as before. Our covariates will continue to be sampled from an R-sub-Gaussian distribution $\mathcal{D}$ with support over $\mathbb{R}^d$. Even in this setting, we can ensure a model recovery result with a linear rate of convergence.

Theorem 2. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted with the rest subjected to sub-Gaussian noise sampled from a distribution $\mathcal{D}_\varepsilon$ with sub-Gaussian norm σ. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$, where $c_\varepsilon > 0$ is a constant that depends only on the distributions $\mathcal{D}$ and $\mathcal{D}_\varepsilon$, then with probability at least $1 - \exp(-\tilde{\Omega}(n))$, after $K = \mathcal{O}\left(\log \frac{1}{M_1 \sigma}\right)$ stages, each of which has only $\mathcal{O}(1)$ iterations, we must have $\|\mathbf{w}^K - \mathbf{w}^*\|_2 \leq \mathcal{O}(\sigma)$.*

We refer the reader to Appendix E for the full proof.

Global Convergence This result also allows arbitrary initialization so long as we set $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$. However, note that this result only guarantees a convergence to $\|\mathbf{w}^{K,1} - \mathbf{w}^*\|_2 \leq \mathcal{O}(\sigma)$ and thus, does not ensure a consistent solution. We refer the reader to the proof of Theorem 2 in Appendix E for a discussion on this result. We also note that our results or our algorithms, do not require the knowledge of the noise parameter σ.

Breakdown Point For Gaussian covariates i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_d)$, Gaussian noise i.e. $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$, we have $c \geq 0.52$ (see Appendix E), and for $\eta \to 1$ this gives STIR and STIR-GD with a breakdown point of $\frac{1}{12.25}$.

8 Experiments

In this section, we report results of a variety of experiments comparing STIR and STIR-GD to other robust learning algorithms. These experiments were performed over two learning settings, namely robust linear regression and robust linear-armed bandit problems.

Parameter and Adversary Setting Algorithms considered in this section require only scalar parameters to be specified (α for TORRENT, step length for TORRENT-GD, η and M_1 for STIR, and step length C for STIR-GD), all which were tuned via a fine grid search using a held-out validation set. In particular, a binary search was found to suffice for setting M_1. For all experiments, the adversary was made to introduce corruptions using a fake model as described in §5. All algorithms were initialized at the fake model itself to test their behavior under adversarial initialization.

8.1 Robust Regression Experiments

We executed STIR and STIR-GD on linear regression problems with response corruption as described in §4.

---

[2]We can tolerate noise with non-zero mean as well, by using a simple pairing trick which has a side effect of at most doubling the corruption rate α

(a) STIR vs IRLS with fixed M

(b) STIR vs TORRENT
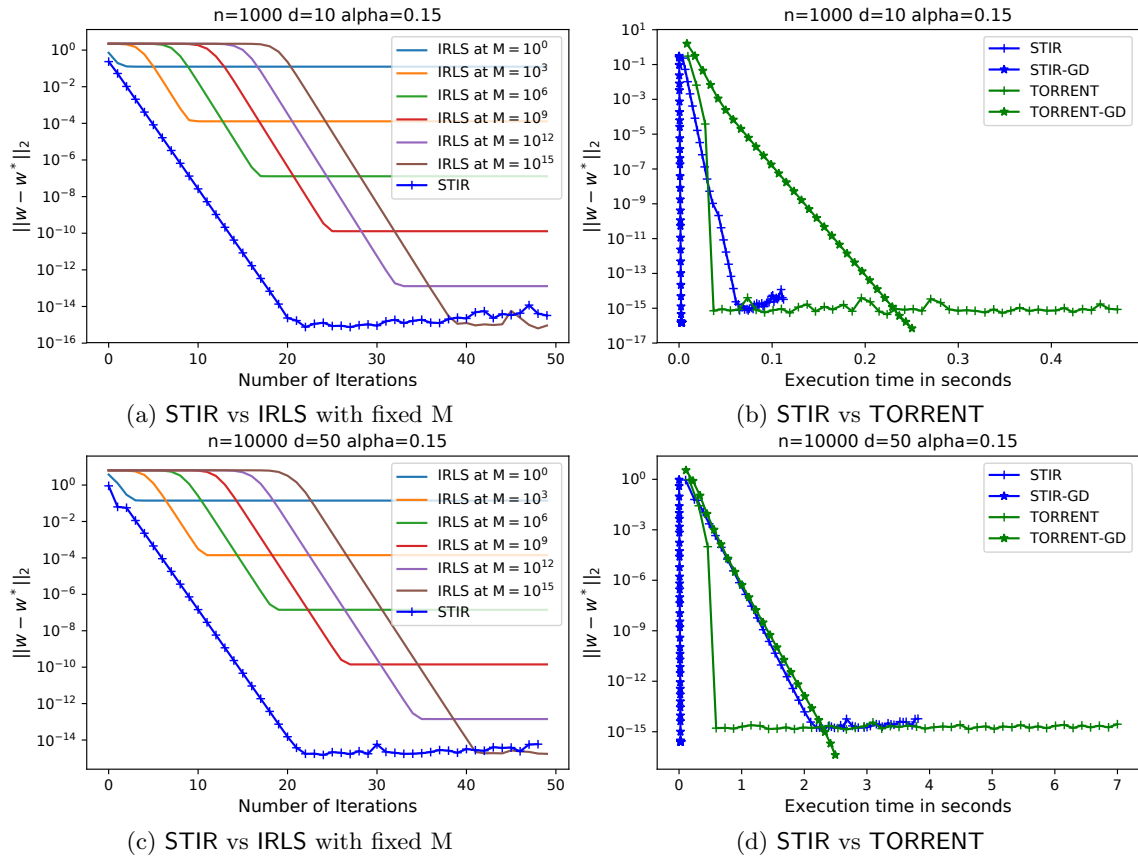
(c) STIR vs IRLS with fixed M

(d) STIR vs TORRENT

Figure 3: All y-axes are in log-scale. Figs (a) and (c) use different data dimensionalities and number of data points and compare STIR to when IRLS is executed with various fixed values of the truncation parameter M. It is clear that no fixed value performs well. For small fixed values $M \approx 10^0$, IRLS converges rapidly but to poor models. For large fixed values $M \approx 10^{12}$, IRLS gets stuck at the fake model and takes long to converge. On the other hand, although STIR was initialized with $M_1 = 0$ for this experiment, it adaptively increases its truncation parameter to offer far better convergence than IRLS with any fixed value of M. Figs (b) and (d) compare STIR and STIR-GD with TORRENT and TORRENT-GD. In all cases, STIR-GD offers the fastest convergence.

(a) Variation with dataset size

(b) Variation with dimension

(c) Variation with corruption
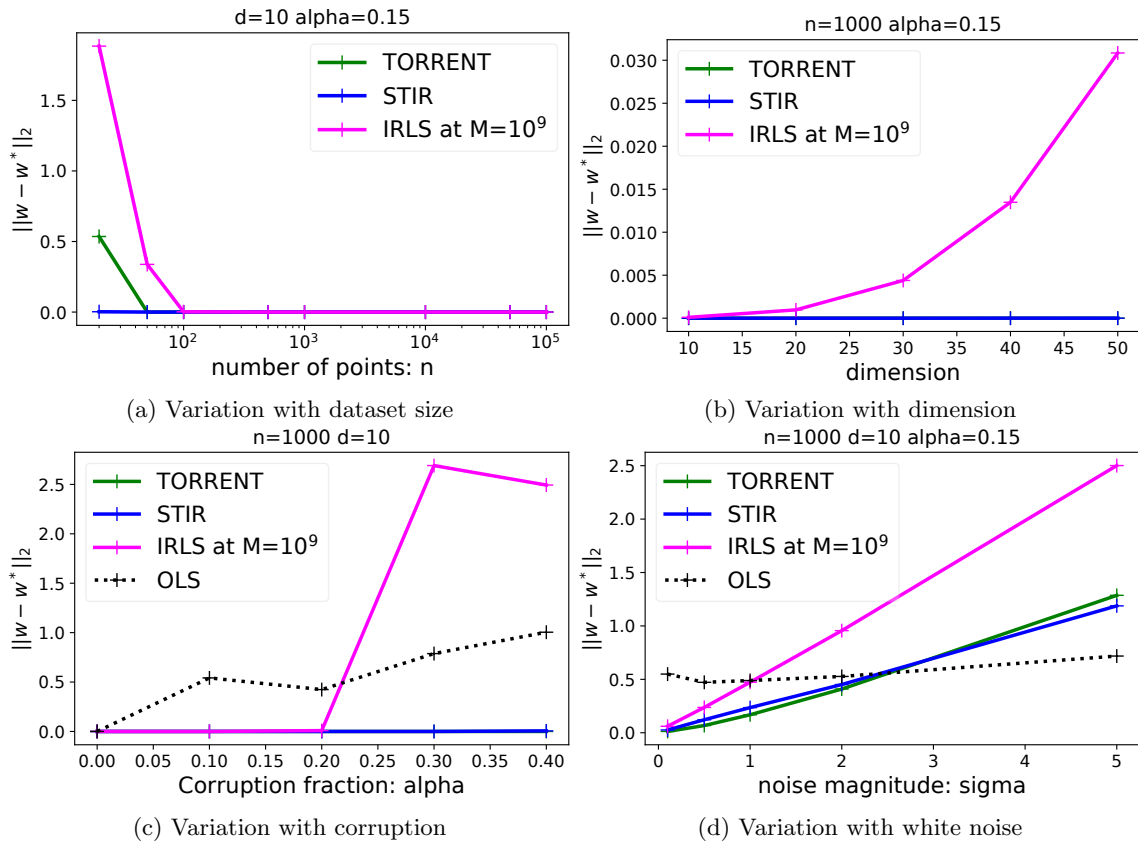
(d) Variation with white noise

Figure 4: The figures compare STIR, TORRENT, IRLS, and OLS for convergence behavior. OLS exceeds the figure boundaries and hence not visible in Figs (a) and (b). Fig (a) examines the effect of varying the training set size. Note that the x-axis is in log-scale. IRLS performs poorly with very few data points but STIR and TORRENT continue to offer good convergence. Fig (b) shows that IRLS worsens with increasing dimensionality whereas STIR and TORRENT remain stable. Fig (c) explores the affect of increasing the fraction of corrupted points. Both OLS and IRLS show considerable worsening with increasing fraction of corruptions. Finally, Fig (d) explores the hybrid noise model discussed in Section 7.3 (Figs (a)-(c) had no white noise). Here, IRLS performs the worst of all. However, once the noise variance goes beyond a point, TORRENT and STIR start losing the distinction between good and bad points and the naive OLS starts outperforming them.
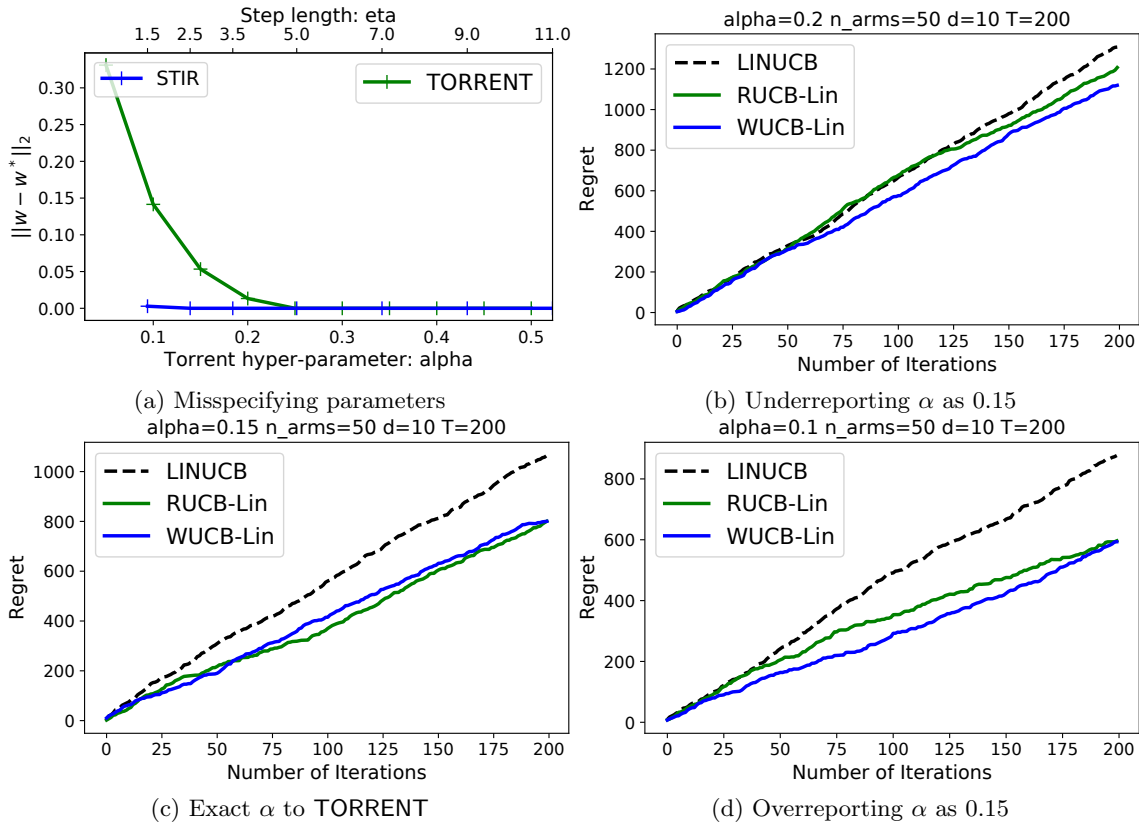
Figure 5: The figures compare STIR and TORRENT with respect to hyperparameter misspecification. STIR was initialized at $\mathbf{w}^0 = \mathbf{0}$ in these experiments. For Fig (a), 25% data was corrupted but TORRENT was given various values of its hyperparameter α (denoting the fraction of corrupted points) as indicated. STIR was also given various values of its own hyperparameter η in a wide range. TORRENT is very susceptible to hyperparameter misspecification and degrades heavily when not given a proper value whereas STIR is much more stable with respect to its hyperparameter. For Figs (b), (c), (d), respectively 20%, 15% and 10% of the data was corrupted and linear-armed bandit algorithms that use OLS (LINUCB), TORRENT (RUCB-Lin) and STIR (WUCB-Lin) were executed. For Figs (b), (c), (d), TORRENT was always given a hyperparameter value $\alpha = 0.15$. Note that this is appropriate for Fig (c) where actually 15% data was corrupted but not for Figs (b) and (d). TORRENT performs comparably to STIR if provided the true value of α, as in Fig (c) but its performance degrades if we give a value smaller than true value, such as in Fig (b) or a larger value, such as in Fig (d).

Algorithm 3 WUCB-Lin: Weighted UCB for Linear Contextual Bandits

---

Input: Upper bounds σ_0 (on sub-Gaussian norm of noise distribution), B (on magnitude of corruption), α_0 (on fraction of corrupted points), initial truncation M_1, increment rate η

1: **for** $t = 1, 2, \ldots, T$ **do**
2:     Receive set of arms A_t
3:     Play arm $\hat{\mathbf{x}}^t = \underset{\mathbf{x} \in A_t, \mathbf{w} \in C_{t-1}}{\arg\max} \; \langle \mathbf{x}, \mathbf{w} \rangle$
4:     Receive reward r_t
5:     $(\hat{\mathbf{w}}^t, S^t) \leftarrow \mathsf{STIR}\left(\{\hat{\mathbf{x}}^\tau, r_\tau\}_{\tau=1}^t, M_1, \eta\right)$
                                                              `//Denote` $S^t = \mathtt{diag}(s_1^t, s_2^t, \ldots, s_t^t)$
6:     $V^t \leftarrow \sum_{\tau \le t} s_\tau^t \hat{\mathbf{x}}^\tau (\hat{\mathbf{x}}^\tau)^\top$, $X^t \leftarrow \left[\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots, \hat{\mathbf{x}}^t\right]$
7:     $\bar{\mathbf{w}}^t \leftarrow (V^t)^{-1} X^t S^t \mathbf{y}$
8:     $C_t \leftarrow \{\mathbf{w} : \|\mathbf{w} - \bar{\mathbf{w}}^t\|_{V^t} \le \sigma_0 \sqrt{d \log T} + \alpha_0 B T\}$
9: **end for**

Algorithms: We compared STIR and $\mathsf{STIR\text{-}GD}$ with the $\mathsf{TORRENT}$ algorithm [6], its faster gradient version $\mathsf{TORRENT\text{-}GD}$, the classical IRLS algorithm with various fixed values of the truncation parameter, and the standard OLS (Ordinary Least Squares) algorithm. We do not compare to some other state-of-the-art algorithms for robust regression, such as L_1 minimization techniques and extended Lasso since [6] establishes that $\mathsf{TORRENT}$ outperforms all of them.

Data: The covariate dimensionality and the number of data points are mentioned with each plot. All covariates were generated from a normal distribution. The gold and fake models were chosen as two independently sampled unit vectors. The set of "bad" data points was chosen randomly and the fake model was used to introduce corruptions, as in Section 5.

8.2   Robust Linear Bandit Experiments

As linear-armed bandit algorithms [1] utilize regression routines internally, recent works have explored the possibility of using robust regression algorithms to target cases when arm-pulls are corrupted, for example [19] that uses $\mathsf{TORRENT}$ itself to develop corruption-tolerant bandit learning algorithms.

Algorithm 3 presents $\mathsf{WUCB\text{-}Lin}$, an adaptation of STIR to linear bandit settings. We refer the reader to Appendix F for details of the algorithm. $\mathsf{WUCB\text{-}Lin}$ roughly follows the popular *Optimism-in-the-face-of-uncertainty* (OFUL) principle while selecting arms to pull at various time instants.

However, since we know some of the arm pulls generated corrupted rewards, instead of applying the OFUL principle blindly, $\mathsf{WUCB\text{-}Lin}$ invokes STIR and obtains not only an estimate of the reward generating model, but also a set of weights on previous arm pulls which indicate which pulls were corrupted and which pulls were clean. $\mathsf{WUCB\text{-}Lin}$ then uses these weights to form a *weighted confidence set* (Algorithm 3, line 6) that is further utilized in applying the OFUL principle to decide future arm pulls (Algorithm 3, line 3).

Algorithms and Data: We compare $\mathsf{WUCB\text{-}Lin}$ with LINUCB that uses the simple OLS estimator, as well as the $\mathsf{RUCB\text{-}Lin}$ algorithm from [19]. We refer the reader to Appendix F for details of the problem setting.

8.3   Discussion on Experiments

Figures 3, 4 and 5 present graphs with the outcomes of the experiments. Although the respective captions in the figures detail the observed behaviours of various algorithms considered therein, here we point out some broad inferences.

1. $\mathsf{STIR\text{-}GD}$ offers much faster convergence as compared to $\mathsf{TORRENT}$ or $\mathsf{TORRENT\text{-}GD}$.

2. No single value of the truncation parameter M ensures a good performance with IRLS. A stage-wise implementation with continuously updated truncation parameters, as STIR offers, is necessary for rapid and assuredly global convergence.

3. TORRENT requires an estimate of the fraction of corrupted points as a hyperparameter and is extremely susceptible to misspecification in this value. STIR and STIR-GD on the other hand are much more resilient to misspecifications of their own hyperparameters.

9 Conclusion and Future Work

In this work we presented STIR, a stage-wise algorithm that makes simple and efficient modifications, including a gradient-based implementation STIR-GD, to the well-known IRLS heuristic to obtain the first global convergence results for robust regression. These algorithms offer not only theoretically superior results to state-of-the-art algorithms such as TORRENT but are empirically faster and more immune to hyperparameter mis-specification.

Our theoretical results are superior to those of previous works in terms of offering a better breakdown point, and are based on a novel notion of weighted strong convexity. Working with this new notion of strong convexity required us to develop the peeling proof technique which is novel in robust regression literature and may be of independent interest in analyzing other iterative algorithms.

Several avenues of future work exist. It would be interesting to examine other weighing functions (IRLS and STIR use the inverse of the residual) for robust regression. It is likely that any reasonable decreasing function of residuals should suffice. It would also be interesting to derive formal regret bounds for the WUCB-Lin algorithm and see how they compare to the regret bounds of the RUCB-Lin algorithm from [19].

Acknowledgements

References

[1] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved Algorithms for Linear Stochastic Bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

[2] Khurrum Aftab and Richard Hartley. Convergence of Iteratively Re-weighted Least Squares to Robust M-Estimators. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[3] Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

[4] Demba Ba, Behtash Babadi, Patrick L. Purdon, and Emery N. Brown. Convergence and Stability of Iteratively Re-weighted Least Squares Algorithms. *IEEE Transactions on Signal Processing*, 62(1):183–195, 2013.

[5] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent Robust Regression. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.

[6] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust Regression via Hard Thresholding. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[7] Nicolai Bissantz, Lutz Dümbgen, Axel Munk, and Bernd Stratmann. Convergence Analysis of Generalized Iteratively Reweighted Least Squares Algorithms on Convex Function Spaces. *SIAM Journal of Optimization*, 19(4):1828–1845, 2009.

[8] Jack Brimberg and Robert F. Love. Global Convergence of a Generalized Iterative Procedure for the Minisum Location Problem with lp Distances. *Operations Research*, 41(6):1010–1176, 1993.

[9] Emmanuel J. Candès, Xiaodong Li, and John Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(1):1–37, 2009.

[10] Yin Chen and Arnak S. Dalalyan. Fused sparsity and robust estimation for linear models with unknown variance. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

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

[12] A. K. Cline. Rate of Convergence of Lawson's Algorithm. *Mathematics of Computation*, 26(117):167–176, 1972.

[13] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C. Sinan Güntürk. Iteratively Reweighted Least Squares Minimization for Sparse Recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[14] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly Learning a Gaussian: Getting Optimal Error, Efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2683–2702, 2018.

[15] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient Algorithms and Lower Bounds for Robust Linear Regression. In *30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

[16] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust Logistic Regression and Classification. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[17] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making Machine Learning Robust Against Adversarial Inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[18] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[19] Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. Machine Learning (to appear) https://doi.org/10.1007/s10994-018-5758-5, 2018.

[20] Lihong Li, Wei Chu, John Langford, and Robert Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International World Wide Web Conference (WWW)*, 2010.

[21] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High Dimensional Robust Sparse Regression. arXiv:1805.11643v1 [cs.LG], 2018.

[22] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 114–122, 2018.

[23] Julien Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM Journal of Optimization*, 25(2):829–855, 2015.

[24] 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.

[25] M. R. Osborne. *Finite Algorithms in Optimization and Data Analysis.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, 1985.

[26] Damian Straszak and Nisheeth K. Vishnoi. IRLS and Slime Mold: Equivalence and Convergence. arXiv:1601.02712 [cs.DS], 2016.

[27] John W. Tukey. A Survey of Sampling from Contaminated Distributions. *Contributions to Probability and Statistics*, 2:448–485, 1960.

[28] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

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

A   IRLS and the Scaled Huber Loss - Supplementary Details

We recapitulate below the definitions of the Huber loss, the scaled (and translated) Huber loss and, given a model $\mathbf{w}^0$ and data $(\mathbf{x}_i, y_i)_{i=1}^n$, other allied functions.

$$h_\epsilon(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq \epsilon \\ \epsilon\,|x| - \frac{1}{2}\epsilon^2 & |x| > \epsilon \end{cases}$$

$$f_\epsilon(x) = \begin{cases} \frac{1}{2}\left(\frac{x^2}{\epsilon} + \epsilon\right) & |x| \leq \epsilon \\ |x| & |x| > \epsilon \end{cases}$$

$$g_\epsilon(x; a) := \frac{1}{2}\left(\frac{x^2}{\max\{|a|, \epsilon\}} + \max\{|a|, \epsilon\}\right)$$

$$\ell_\epsilon(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^n f_\epsilon\left(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i\right)$$

$$\wp_\epsilon(\mathbf{w}; \mathbf{w}^0) := \sum_{i=1}^n g_\epsilon\left(\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i; \langle \mathbf{w}^0, \mathbf{x}_i \rangle - y_i\right)$$

The claim that M-truncated IRLS minimizes $\wp_{\frac{1}{M}}(\mathbf{w}; \mathbf{w}^0)$ to obtain the next model can be easily verified using the equivalence between the truncation and regularization techniques explained in Footnote 1 (see §5 for the footnote). In the following, we establish that $g_\epsilon(\cdot; \cdot)$ is a valid majorizer for f_ϵ for any $\epsilon > 0$.

Claim 3. *For any $a, x \in \mathbb{R}, \epsilon > 0$, we have $g_\epsilon(a; a) = f_\epsilon(a)$ as well as $g_\epsilon(x; a) \geq f_\epsilon(x)$.*

Proof. We have, for the first claim,

$$g_\epsilon(a; a) = \frac{1}{2}\left(\frac{a^2}{\max\{|a|, \epsilon\}} + \max\{|a|, \epsilon\}\right) = \begin{cases} \frac{1}{2}\left(\frac{a^2}{\epsilon} + \epsilon\right) & |a| \leq \epsilon \\ |a| & |a| > \epsilon \end{cases} = f_\epsilon(a).$$

For the second claim, we consider two simple cases

Case 1 $|x| > \epsilon$: In this case we have $f_\epsilon(x) = |x|$ and we always have $\frac{1}{2}\left(\frac{x^2}{\max\{|a|,\epsilon\}} + \max\{|a|, \epsilon\}\right) \geq |x|$.

Case 2 $|x| \leq \epsilon$: In this case denote $b = \max\{|a|, \epsilon\}$. Then we have $b \geq \epsilon \geq |x|$ which gives us $x^2 \leq b\epsilon$. Thus, we have $g_\epsilon(x; a) - f_\epsilon(x) = \frac{1}{2}\left(\frac{x^2}{b} + b\right) - \frac{1}{2}\left(\frac{x^2}{\epsilon} + \epsilon\right) = \frac{(b-\epsilon)(b\epsilon - x^2)}{2b\epsilon} \geq 0$. □

The following claim shows that we have $f'_\epsilon(x)|_{x=a} = g'_\epsilon(x; a)|_{x=a}$ for any ϵ, a. This immediately establishes that $\nabla\wp_\epsilon(\mathbf{w}^0; \mathbf{w}^0) = \nabla\ell_\epsilon(\mathbf{w}^0)$ for any model $\mathbf{w}^0$.

Claim 4. *For any $a, x \in \mathbb{R}, \epsilon > 0$, we have $f'_\epsilon(x)|_{x=a} = g'_\epsilon(x; a)|_{x=a}$.*

Proof. We have $g'_\epsilon(x; a) = \frac{x}{\max\{|a|, \epsilon\}}$ which gives us

$$g'_\epsilon(x; a)|_{x=a} = \begin{cases} \frac{a}{\epsilon} & |a| \leq \epsilon \\ \text{sign}(a) & |a| > \epsilon, \end{cases}$$

whereas we have

$$f'_\epsilon(x) = \begin{cases} \frac{x}{\epsilon} & |x| \leq \epsilon \\ \text{sign}(x) & |x| > \epsilon \end{cases},$$

which establishes the claim. $\qquad\square$

B   Supporting Results

In this section we prove a few results used in the convergence analysis of STIR.

Lemma 5. *Suppose we have data covariates $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ generated from an isotropic but otherwise arbitrary sub-Gaussian distribution. Then for any fixed set $S \subset [n]$ and $n = \Omega\left(d + \log \frac{1}{\delta}\right)$, with probability at least $1 - \delta$,*
$$0.99 |S| \leq \lambda_{\min}(X_S X_S^\top) \leq \lambda_{\max}(X_S X_S^\top) \leq 1.01 |S|,$$
where the constant inside $\Omega(\cdot)$ depends only on the sub-Gaussian distribution and universal constants.

Proof. This is a special case of [6, Lemma 16] for isotropic distributions. Note that since our adversary is partially adaptive, the sets of good and bad points G, B are fixed and this lemma applies to both G and B. $\qquad\square$

Lemma 6. *Suppose our data covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are generated from a sub-Gaussian distribution with sub-Gaussian norm R. Then with probability at least $1 - \delta$, we have $R_X := \max_{i \in [n]} \|\mathbf{x}_i\|_2 \leq \|\boldsymbol{\mu}\|_2 + \mathcal{O}\left(R\sqrt{d + \log \frac{n}{\delta}}\right)$.*

Proof. If $\mathbf{x}$ is R-sub-Gaussian with mean $\boldsymbol{\mu}$, then for any unit vector $\mathbf{v} \in S^{d-1}$, $\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu} \rangle$ is centered as well as $2R$-sub-Gaussian which gives us

$$\mathbb{P}\left[|\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu} \rangle| \geq t\right] \leq 2 \exp\left[-t^2/2R^2\right]$$

If $\mathbf{v}^1, \mathbf{v}^2 \in S^{d-1}$, such that $\|\mathbf{v}^1 - \mathbf{v}^2\|_2 \leq \frac{1}{2}$, then we have $|\langle \mathbf{v}^1 - \mathbf{v}^2, \mathbf{x} - \boldsymbol{\mu} \rangle| \leq \frac{1}{2} \cdot \|\mathbf{x} - \boldsymbol{\mu}\|_2$. Thus, taking a union bound over a $1/2$-net over S^{d-1} gives us

$$\mathbb{P}\left[\max_{\mathbf{v} \in S^{d-1}} |\langle \mathbf{v}, \mathbf{x} - \boldsymbol{\mu} \rangle| \geq \frac{1}{2} \cdot \|\mathbf{x} - \boldsymbol{\mu}\|_2 + t\right] = \mathbb{P}\left[\|\mathbf{x}\|_2 \geq \|\boldsymbol{\mu}\|_2 + 2t\right] \leq 2 \cdot 5^d \exp\left[-t^2/2R^2\right]$$

Taking $t^2 = 2R^2(d \log 5 + \log \frac{n}{\delta} + \log 2)$ proves the result.

$$\mathbb{P}\left[\max_{i \in [n]} \|\mathbf{x}_i\|_2 > \|\boldsymbol{\mu}\|_2 + R\sqrt{2\left(d \log 5 + \log \frac{n}{\delta} + \log 2\right)}\right] \leq \delta \qquad\square$$

In the following, we establish that the scaled Huber loss is Lipschitz. This will be helpful in transferring our convergence guarantees to those with respect to the Huber and absolute loss functions.

Lemma 7. *For any $\epsilon > 0$, we have $|\ell_\epsilon(\mathbf{w}) - \ell_\epsilon(\mathbf{w}')| \leq \|\mathbf{w} - \mathbf{w}'\|_2 \cdot \sqrt{1.01}$.*

Proof. The function $f_\epsilon(\cdot)$ is clearly 1-Lipschitz for any $\epsilon > 0$. This means that we have

$$|\ell_\epsilon(\mathbf{w}) - \ell_\epsilon(\mathbf{w}')| \leq \frac{1}{n} \sum_{i=1}^{n} |\langle \mathbf{w}, \mathbf{x}_i \rangle - \langle \mathbf{w}', \mathbf{x}_i \rangle| = \frac{1}{n} \left\| X^\top (\mathbf{w} - \mathbf{w}') \right\|_1 \leq \frac{1}{\sqrt{n}} \left\| X^\top (\mathbf{w} - \mathbf{w}') \right\|_2$$

$$\leq \frac{1}{\sqrt{n}} \left\| X \right\|_2 \left\| \mathbf{w} - \mathbf{w}' \right\|_2 \leq \left\| \mathbf{w} - \mathbf{w}' \right\|_2 \cdot \sqrt{1.01},$$

where the last step follows due to Lemma 5. $\qquad\square$

C   Convergence Analysis - Supplementary Details

We begin by restating Theorem 1, the main result that we will prove in this section.

Theorem 1. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \leq \frac{c}{2.88\eta + c}$, where $c > 0$ is a constant that depends only on $\mathcal{D}$, then for any $\epsilon > 0$, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d + n) + \log\frac{1}{M_1\epsilon}\right)\right)$, after $K = \mathcal{O}\left(\log\frac{1}{M_1\epsilon}\right)$ stages, we must have $\left\|\mathbf{w}^K - \mathbf{w}^*\right\|_2 \leq \epsilon$. Moreover, each stage consists of only $\mathcal{O}(1)$ iterations.*

Proof. As mentioned before, notice that this is indeed a global convergence guarantee since it places no restrictions on the initial model $\mathbf{w}^0$. The only requirement is that the accompanying initial truncation parameter M_1 complement the model initialization by satisfying $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$. In particular, if initialized at the origin, as Algorithms 1 and 2 do, we need only ensure $M_1 \leq \frac{1}{R_W}$ where $R_W = \|\mathbf{w}^*\|_2$. This can be done using a simple binary search to identify an appropriate value of M_1. Recall that both STIR and STIR-GD operate in stages. We introduce a notion of a *well-initialized* stage below.

Definition 2 (Well-initialized Stage)**.** *A stage in the execution of STIR or STIR-GD is said to be well-initialized if, given the truncation parameter M_T which will be used during that stage, at the beginning of that stage T, we are in possession of a model $\mathbf{w}^{T,1}$ that satisfies $\left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M_T}$.*

Note that the initialization of STIR and STIR-GD with respect to the setting of M_1 ensure $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$ which implies that the very first stage is always well-initialized. Now, Lemmata 8 and 9 show that, if the preconditions of this theorem are satisfied, then a stage T, started off with a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ (see Algorithm 1, line 3) and a truncation parameter M_T that satisfy the well-initialized condition i.e. $\left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M_T}$, will ensure with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d + n)\right)\right)$, that there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M_T}$ for all $\tau \geq t_0$.

An application of the triangle inequality shows that we will have $\left\|\mathbf{w}^{T,t_0} - \mathbf{w}^{T,t_0+1}\right\|_2 \leq \frac{2}{\eta M_T}$ which implies (see Algorithm 1, line 5) that we will exit this stage at the $(t_0 + 1)^{\text{th}}$ inner iteration. However, notice that at this point we are endowed with $\left\|\mathbf{w}^{T+1,1} - \mathbf{w}^*\right\|_2 = \left\|\mathbf{w}^{T+1} - \mathbf{w}^*\right\|_2 = \left\|\mathbf{w}^{T,t_0+1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M_T} = \frac{1}{M_{T+1}}$. Note that this means that stage $(T + 1)$ is well-initialized too.

Thus, whenever a stage T is well-initialized, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d + n)\right)\right)$, we have $\left\|\mathbf{w}^{T+1,1} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta} \left\|\mathbf{w}^{T,1} - \mathbf{w}^*\right\|_2$. Since we always set $\eta > 1$, there exists an upper bound $T_0 = \mathcal{O}\left(\log\frac{1}{M_1\epsilon}\right)$ on the number of stages. Thus, an application of union bound shows that we must have $\left\|\mathbf{w}^{T_0+1,1} - \mathbf{w}^*\right\|_2 \leq \epsilon$ with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d + n)\right) + \log\frac{1}{M_1\epsilon}\right) = 1 - \exp(-\tilde{\Omega}(n))$ for all $\epsilon = \frac{1}{n^{\mathcal{O}(1)}}$. $\qquad\square$

Lemma 8. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. Suppose we initialize a stage T within an execution of*

STIR with truncation level M, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ such that $\alpha \leq \frac{c}{2.88\eta + c}$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$. Here c is the constant of the WSC property and depends only on the distribution $\mathcal{D}$ (see Lemma 12).

Proof. Let $\mathbf{w}^{T,\tau}$ be a model encountered by STIR within this stage and let $\mathbf{r} = X^\top \mathbf{w}^{T,\tau} - \mathbf{y}$ denote the residuals due to $\mathbf{w}^{T,\tau}$ and $S = \mathrm{diag}(\mathbf{s})$ denote the diagonal matrix of weights where $\mathbf{s}_i = \min\left\{\frac{1}{|\mathbf{r}_i|}, M\right\}$. Then STIR will choose as the next model $\mathbf{w}^{T,\tau+1} = (XSX^\top)^{-1}XS\mathbf{y} = \mathbf{w}^* + (XSX^\top)^{-1}XS\mathbf{b}$ which gives us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{\|XS\mathbf{b}\|_2}{\lambda_{\min}(XSX^\top)}$$

Now by Lemma 5, with probability at least $1 - \exp(-\Omega(n - d))$, we have $\|X_B\|_2 = \sqrt{\lambda_{\max}(X_B X_B^\top)} \leq \sqrt{1.01B}$. By Lemma 10 we have, again with probability at least $1 - \exp(-\Omega(n - d))$

$$\|S\mathbf{b}\|_2 \leq \sqrt{4B(1 + 1.01M^2\|\mathbf{w} - \mathbf{w}^*\|_2^2)} \leq 2\sqrt{2.01B}$$

It should be noted that Lemma 10 relies precisely on Lemma 5 to derive its confidence assurance. Since the nature of Lemma 5 is such that it need be established only once, and not repeatedly for every iteration, we have, with probability at least $1 - \exp(-\Omega(n - d))$, for *all iterations* within this stage (actually all iterations across all stages), both Lemma 10 and Lemma 5 hold simultaneously.

Using Lemma 12, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, we have $\lambda_{\min}(XSX^\top) \geq \lambda_{\min}(X_G S_G X_G^\top) \geq 0.99c \cdot GM$. Note that since all models $\mathbf{w}^{T,\tau}, \tau \geq 1$ in this stage will at least satisfy $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M}$ (since the initial model $\mathbf{w}^{T,1}$ satisfies this by assumption and STIR offers monotonic convergence), the result of Lemma 12 applies uniformly to all these models and need not be applied separately to each model in this stage. Using these results to upper bound $\|XS\mathbf{b}\|_2$ and lower bound $\lambda_{\min}(XSX^\top)$ shows that at either we must have

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{2B\sqrt{2.0301}}{0.99c \cdot GM}$$

or else if the above is not true, then we must instead have

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq 0.99 \cdot \|\mathbf{w} - \mathbf{w}^*\|_2$$

Note that since we have $\alpha \leq \frac{c}{2.88\eta + c}$, we get $\frac{2B\sqrt{2.0301}}{0.99c \cdot GM} \leq \frac{1}{\eta M}$. Thus, it is assured that after $t_0 = \mathcal{O}(\log\eta) = \mathcal{O}(1)$ iterations, iterates $\mathbf{w}^{T,\tau}$ of STIR will satisfy $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$ $\qquad\square$

Lemma 9. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted. Suppose we initialize a stage T within an execution of* STIR-GD *with truncation level M, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^{T,1}$ such that $\alpha \leq \frac{c}{2.88\eta + c}$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, there exists an upper bound of $t_0 = \mathcal{O}(1)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$.*

Proof. As observed before, all models $\mathbf{w}^{T,\tau}, \tau \geq 1$ in this stage at least satisfy $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{M}$ since the initial model $\mathbf{w}^{T,1}$ satisfies this by assumption and we will see below that STIR-GD offers monotonic convergence. Thus, Lemma 12 applies uniformly to all these models and thus, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, for all $\tau \geq 1$, the function $\wp_{\frac{1}{M}}(\cdot, \mathbf{w}^{T,\tau})$ (refer to §6 for notation) is γ-strongly convex for $\gamma \geq 0.99c \cdot GM$.

Similarly, Lemma 5 tells us that, again with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, for all $\tau \geq 1$, the function $\wp_{\frac{1}{M}}(\cdot, \mathbf{w}^{T,\tau})$ is δ-strongly smooth for $\delta \leq 1.01Mn$. From now on, we will be using the shorthand $\wp(\cdot) := \wp_{\frac{1}{M}}(\cdot, \mathbf{w}^{T,\tau})$ to avoid notational clutter.

If we denote $\mathbf{g}^t := \nabla \wp(\mathbf{w}^{T,\tau}) = \wp_{\frac{1}{M}}(\mathbf{w}^{T,\tau}, \mathbf{w}^{T,\tau})$, then it is clear that STIR-GD will choose as the next model as $\mathbf{w}^{T,\tau+1} := \mathbf{w}^{T,\tau} - \frac{C}{Mn} \cdot \mathbf{g}^t$. For sake of notational simplicity, we will abbreviate $\mathbf{w} := \mathbf{w}^{T,\tau}, \mathbf{w}^+ := \mathbf{w}^{T,\tau+1}, \mathbf{g} := \mathbf{g}^t$. Then, applying strong smoothness tells us that

$$\wp(\mathbf{w}^+) - \wp(\mathbf{w}) \leq \langle \mathbf{g}, \mathbf{w}^+ - \mathbf{w} \rangle + \frac{\delta}{2} \left\| \mathbf{w}^+ - \mathbf{w} \right\|_2^2$$

$$= \langle \mathbf{g}, \mathbf{w}^+ - \mathbf{w}^* \rangle + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w} \rangle + \frac{\delta}{2} \left\| \mathbf{w}^+ - \mathbf{w} \right\|_2^2$$

$$= \frac{Mn}{C} \cdot \langle \mathbf{w} - \mathbf{w}^+, \mathbf{w}^+ - \mathbf{w}^* \rangle + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w} \rangle + \frac{\delta}{2} \left\| \mathbf{w}^+ - \mathbf{w} \right\|_2^2$$

$$= \frac{Mn}{2C} \left(\left\| \mathbf{w} - \mathbf{w}^* \right\|_2^2 - \left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2^2 \right) + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w} \rangle + \left(\frac{\delta}{2} - \frac{Mn}{2C} \right) \left\| \mathbf{w}^+ - \mathbf{w} \right\|_2^2$$

$$\leq \frac{Mn}{2C} \left(\left\| \mathbf{w} - \mathbf{w}^* \right\|_2^2 - \left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2^2 \right) + \langle \mathbf{g}, \mathbf{w}^* - \mathbf{w} \rangle,$$

where the fifth step holds for any $C \leq \frac{Mn}{\delta} \leq 0.99$. Strong smoothness on the other hand tells us that

$$\langle \mathbf{g}, \mathbf{w}^* - \mathbf{w} \rangle \leq \wp(\mathbf{w}^*) - \wp(\mathbf{w}) - \frac{\gamma}{2} \left\| \mathbf{w} - \mathbf{w}^* \right\|_2^2$$

Combining the above two results gives us

$$\wp(\mathbf{w}^+) - \wp(\mathbf{w}^*) \leq \frac{Mn}{2C} \left(\left\| \mathbf{w} - \mathbf{w}^* \right\|_2^2 - \left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2^2 \right) - \frac{\gamma}{2} \left\| \mathbf{w} - \mathbf{w}^* \right\|_2^2$$

Now, we can either have $\wp(\mathbf{w}^+) - \wp(\mathbf{w}^*) \geq 0$ in which case we get $\left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2 \leq \sqrt{1 - \frac{C\gamma}{Mn}} \left\| \mathbf{w} - \mathbf{w}^* \right\|_2 \leq \sqrt{1 - \frac{0.99cCG}{n}} \left\| \mathbf{w} - \mathbf{w}^* \right\|_2$ or else $\wp(\mathbf{w}^+) - \wp(\mathbf{w}^*) < 0$ in which case applying strong convexity once again yields

$$\frac{\gamma}{2} \left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2^2 \leq \wp(\mathbf{w}^+) - \wp(\mathbf{w}^*) + \langle \nabla \wp(\mathbf{w}^*), \mathbf{w}^* - \mathbf{w}^+ \rangle \leq \langle \nabla \wp(\mathbf{w}^*), \mathbf{w}^* - \mathbf{w}^+ \rangle$$

Now notice that $\nabla \wp(\mathbf{w}^*) = XS\mathbf{b}$ and Lemmata 10 and 5 tell us that $\left\| XS\mathbf{b} \right\|_2 \leq 2B\sqrt{5.05}$ which give us $\left\| \mathbf{w}^+ - \mathbf{w}^* \right\|_2 \leq \frac{2B\sqrt{2.0301}}{\gamma} \leq \frac{2B\sqrt{2.0301}}{0.99cGM} < \frac{1}{\eta M}$ whenever $\frac{B}{G} \leq \frac{0.99c}{2\eta\sqrt{2.0301}}$. This completes the proof of the result upon making similar arguments as those made in the proof of Lemma 9. $\square$

C.1 Bounding the Weights on Bad Points

The following lemma establishes that neither STIR nor STIR-GD put too much weight on bad points.

Lemma 10. *Suppose during the execution of STIR or STIR-GD, we encounter a model $\mathbf{w}$ while the truncation parameter is M. Denote $\left\| \mathbf{w} - \mathbf{w}^* \right\|_2 = \epsilon$ and let $S = diag(\mathbf{s})$ be the diagonal matrix of M-truncated weights assigned due to residuals induced by $\mathbf{w}$. Then, with probability at least $1 - \exp(-\Omega(n-d))$, we must have*

$$\left\| S\mathbf{b} \right\|_2^2 \leq 4B(1 + 1.01M^2\epsilon^2),$$

where we recall that $\mathbf{b}$ denotes the vector of corruptions.

Proof. Let $\boldsymbol{\Delta} := \mathbf{w} - \mathbf{w}^*$ and let b_i denote the corruption on the data point $\mathbf{x}_i$. The proof proceeds via a simple case analysis

Case 1: $|b_i| \leq 2 |\boldsymbol{\Delta} \cdot \mathbf{x}_i|$ In this case we simply bound $(s_i b_i)^2 \leq M^2 b_i^2 \leq 4M^2 (\boldsymbol{\Delta} \cdot \mathbf{x}_i)^2$.

Case 2: $|b_i| > 2 |\boldsymbol{\Delta} \cdot \mathbf{x}_i|$ In this case we have $|r_i| = |\boldsymbol{\Delta} \cdot \mathbf{x}_i - b_i| \geq |b_i| - |\boldsymbol{\Delta} \cdot \mathbf{x}_i| \geq \frac{|b_i|}{2}$ and thus we must have $s_i \leq \frac{2}{|b_i|}$ (due to possible truncation) and thus $(s_i b_i)^2 \leq 4$.

19

Thus, we get

$$\|S\mathbf{b}\|_2^2 = \sum_{i \in B}(s_i b_i)^2 \leq 4 \cdot \sum_{i \in B} \max\left\{1, M^2(\mathbf{\Delta} \cdot \mathbf{x}_i)^2\right\} \leq 4(B + M^2 \epsilon^2 \lambda_{\max}(X_B X_B^\top)) \leq 4(B + 1.01 M^2 \epsilon^2 B),$$

where the last step follows due to Lemma 5 which holds with probability at least $1 - \exp(-\Omega(n-d))$ and finishes the proof. □

C.2 Convergence with respect to Huber and Absolute Loss

A relatively straightfoward application of Theorem 1 alongwith some Lipschitzness properties allows us to show that STIR and STIR-GD also ensure convergence to the optimal objective value with respect to the Huber and absolute loss functions. These are widely used in robust regression applications.

Theorem 11. *Under the same preconditions as those in Theorem 1, we are assured with probability at least* $1 - \exp(-\tilde{\Omega}(n))$, *that after* $K = \mathcal{O}\left(\log \frac{1}{M_1 \epsilon}\right)$ *stages, both* STIR *and* STIR-GD *must produce a model* $\mathbf{w}^K$ *so that*

 1. $\ell_\epsilon(\mathbf{w}^K) \leq \ell_\epsilon(\mathbf{w}^*) + \sqrt{1.01}\epsilon$

 2. $\frac{1}{n}\left\|X^\top \mathbf{w}^K - \mathbf{y}\right\|_1 \leq \frac{1}{n}\left\|X^\top \mathbf{w}^* - \mathbf{y}\right\|_1 + \frac{3\sqrt{1.01}}{2}\epsilon.$

Proof. The first part follows directly from Lemma 7 and Theorem 1. The second part follows due to the fact that $|x| \leq f_\epsilon(x) \leq |x| + \frac{\epsilon}{2}$ for any $\epsilon > 0$ and thus,

$$\frac{1}{n}\left\|X^\top \mathbf{w}^K - \mathbf{y}\right\|_1 \leq \ell_\epsilon(\mathbf{w}^K) \leq \ell_\epsilon(\mathbf{w}^*) + \sqrt{1.01}\epsilon \leq \frac{1}{n}\left\|X^\top \mathbf{w}^* - \mathbf{y}\right\|_1 + \frac{3\sqrt{1.01}}{2}\epsilon,$$

where the second inequality in the above chain follows from part 1 of this claim. □

D Establishing WSC/WSS - Supplementary Details

Recall that for any $r > 0$ and $M > 0$, $\mathcal{S}_M(r)$ denotes the set of all diagonal M-truncated weight matrices STIR could possibly generate with respect to models residing in the radius R ball centered at $\mathbf{w}^*$ i.e.

$$\mathcal{S}_M(r) := \left\{S = \text{diag}(\mathbf{s}), \mathbf{s}_i = \min\left\{\frac{1}{|\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i|}, M\right\}, \mathbf{w} \in \mathcal{B}_2(\mathbf{w}^*, r)\right\},$$

then we have the following result.

Lemma 12. *Suppose the data covariates* $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ *are generated from an isotropic R-sub-Gaussian distribution* $\mathcal{D}$, *and G denotes the set of uncorrupted points (as well as the size of that set) then there exists a constant c that depends only on the distribution* $\mathcal{D}$ *such that for any fixed value of $M > 0$,*

$$\left.\begin{array}{l}\mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99 c \cdot GM\right] \\ \mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\max}(X_G S_G X_G^\top) > 1.01 \cdot GM\right]\end{array}\right\} \leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right),$$

where the constants inside $\Omega(\cdot)$ are clarified in the proof. In particular, if $\mathcal{D}$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, I_d)$, *then we can take $c = 0.96$.*

Proof. The bound for the largest eigenvalue follows directly due to the fact that all weights are upper bounded by M and hence $X_G S_G X_G^\top \preceq M \cdot X_G X_G^\top$ and applying Lemma 5. For the bound on the smallest

eigenvalue, notice that Lemma 14 shows us that for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$, i.e. a set of M-truncated weights that correspond to some fixed model $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$, we have

$$\mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995c \cdot GM\right] \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right]$$

Recall that we let $R_X := \max_{i \in [n]} \|\mathbf{x}_i\|_2$ denote the maximum Euclidean length of any covariate. However, Lemma 15 shows us that if $\mathbf{w}^1, \mathbf{w}^2 \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ are two models such that $\|\mathbf{w}^1 - \mathbf{w}^2\|_2 \leq \tau$ then, conditioned on the value of R_X, the following holds *almost surely*.

$$\left|\lambda_{\min}(X_G S_G^1 X_G^\top) - \lambda_{\min}(X_G S_G^2 X_G^\top)\right| \leq 2G\tau M^2 R_X^3$$

This prompts us to initiate a uniform convergence argument by setting up a τ-net over $\mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ for $\tau = \frac{c}{400 R_X^3 M}$. Note that such a net has at most $\left(\frac{800 R_X^3}{c}\right)^d$ elements by applying standard covering number bounds for the Euclidean ball [28, Corollary 4.2.13]. Taking a union bound over this net gives us

$$\mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99c \cdot GM\right] \leq 2 \cdot \left(\frac{7200 R_X^3}{c}\right)^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right]$$
$$\leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right),$$

where in the last step we used Lemma 6 to bound $R_X = \mathcal{O}\left(R\sqrt{d+n}\right)$ with probability at least $1 - \exp(-\Omega(n))$. For the specific bound on the constant c for various distributions, including the Gaussian distribution, we refer the reader to Section D.1. $\qquad\square$

The proof of the above result relies on several intermediate results which we prove in succession below. In the first result Lemma 13, we establish expected bounds on the extremal singular values of the matrix $X_G S_G X_G^\top$ corresponding to a fixed model $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$. In the next result Lemma 14, we establish the same result, but this time with high probability instead of in expectation. The next result Lemma 15 establishes that extremal singular values corresponding to two models close to each other must be (deterministically) close.

Lemma 13 (Pointwise Expectation). *With the same preconditions as in Lemma 12, there must exist a constant $c > 0$ that depends only on $\mathcal{D}$ such that for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$, and fixed vector unit $\mathbf{v} \in S^{d-1}$, we have*
$$c \cdot GM \leq \mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] \leq GM.$$
In particular, if $\mathcal{D}$ is the standard Gaussian $\mathcal{N}(\mathbf{0}, I_d)$, then we can take $c = 0.96$.

Proof. Let $\mathbf{x} \sim \mathcal{D}$ and let $y = \langle \mathbf{w}^*, \mathbf{x} \rangle$. Then if we let $\boldsymbol{\Delta} := \frac{\mathbf{w} - \mathbf{w}^*}{\|\mathbf{w} - \mathbf{w}^*\|_2}$ (note that $\|\mathbf{w} - \mathbf{w}^*\| \leq \frac{1}{M}$), then we have $s = \min\left\{\frac{1}{|\langle \mathbf{w}, \mathbf{x} \rangle - y|}, M\right\} \geq M \cdot \min\left\{\frac{1}{|\langle \boldsymbol{\Delta}, \mathbf{x} \rangle|}, 1\right\}$ as well as $s \leq M$. Then by linearity of expectation we have
$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] = \mathbb{E}\left[\sum_{i \in G} \mathbf{s}_i \langle \mathbf{x}_i, \mathbf{v} \rangle^2\right] = G \cdot \mathbb{E}\left[s \cdot \langle \mathbf{x}, \mathbf{v} \rangle^2\right] \leq GM \cdot \mathbb{E}\left[\langle \mathbf{x}, \mathbf{v} \rangle^2\right] = GM,$$

since $\mathcal{D}$ is isotropic. We also get
$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] = G \cdot \mathbb{E}\left[s \cdot \langle \mathbf{x}, \mathbf{v} \rangle^2\right] \geq GM \cdot \mathbb{E}\left[\min\left\{\frac{1}{|\langle \boldsymbol{\Delta}, \mathbf{x} \rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v} \rangle^2\right] \geq c \cdot GM,$$

where, for any distribution $\mathcal{D}$ over $\mathbb{R}^d$, we define the constant c as
$$c := \inf_{\mathbf{u}, \mathbf{v} \in S^{d-1}} \left\{\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\min\left\{\frac{1}{|\langle \mathbf{u}, \mathbf{x} \rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v} \rangle^2\right]\right\}.$$

This concludes the proof. For the specific bound on the constant c for various distributions, including the Gaussian distribution, we refer the reader to Section D.1. $\qquad\square$

Lemma 14 (Pointwise Convergence). *With the same preconditions as in Lemma 12, for any fixed $S \in \mathcal{S}_M(\frac{1}{M})$,*

$$\left.\begin{array}{l} \mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995c \cdot GM\right] \\ \mathbb{P}\left[\lambda_{\max}(X_G S_G X_G^\top) > 1.005 \cdot GM\right] \end{array}\right\} \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005c)^2}{8R^4}\right]$$

Proof. Note that for any square symmetric matrix $A \in \mathbb{R}^{d \times d}$, we have $c - \delta \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq c + \delta$ for some $\delta > 0$ iff $\left|\mathbf{v}^\top A \mathbf{v} - c\right| \leq \delta$ for all $\mathbf{v} \in S^{d-1}$ which itself happens iff $\|A - c \cdot I\|_2 \leq \delta$. Now, if $\mathcal{N}_\epsilon$ denotes an ϵ-net over S^{d-1}, then for any square symmetric matrix $B \in \mathbb{R}^{d \times d}$, we have $\|B\|_2 \leq (1 - 2\epsilon)^{-1} \sup_{\mathbf{v} \in \mathcal{N}_\epsilon} \left|\mathbf{v}^\top B \mathbf{v}\right|$. Thus, setting $B = A - c \cdot I$ and $\epsilon = 1/4$, we have $\|A - c \cdot I\|_2 \leq 2 \sup_{\mathbf{v} \in \mathcal{N}_{1/4}} \left|\mathbf{v}^\top A \mathbf{v} - c\right|$.

Let $\mathbf{x} \sim \mathcal{D}$ and $t = \sqrt{\min\left\{\frac{1}{|\langle \mathbf{w} - \mathbf{w}^*, \mathbf{x}\rangle|}, M\right\}} \leq \sqrt{M}$ and for any fixed $\mathbf{v} \in S^{d-1}$, let $Z := t \cdot \langle \mathbf{x}, \mathbf{v}\rangle$. Then we have

$$\|Z\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}\left[|Z|^p\right]\right)^{1/p} \leq \sqrt{M} \cdot \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}\left[|\langle \mathbf{x}, \mathbf{v}\rangle|^p\right]\right)^{1/p} = R\sqrt{M},$$

where the last step follows by observing that since $\mathcal{D}$ is R-sub-Gaussian, $\|\langle \mathbf{x}_1, \mathbf{v}\rangle\|_{\Psi_2} \leq R$. Thus, Z is $R\sqrt{M}$-sub-Gaussian. This implies Z^2 is MR^2-subexponential (see [28, Lemma 2.7.6]), as well as $Z^2 - \mathbb{E}Z^2$ is $2MR^2$-subexponential by centering and applying the triangle inequality. Note that Lemma 13 implicitly establishes that $\mu := \mathbb{E}Z^2 \in [cM, M]$. Let $Z_1, Z_2, \ldots, Z_G$ be independent realizations of Z with respect to a fixed vector $\mathbf{v}$. Then we have

$$\mathbb{P}\left[\left|\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v} - G\mu\right| \geq \varepsilon \cdot GM\right] = \mathbb{P}\left[\left|\sum_{i \in G}(Z_i^2 - \mu)\right| \geq \varepsilon \cdot GM\right]$$

$$\leq 2\exp\left[-m \cdot \min\left\{\frac{(\varepsilon \cdot GM)^2}{4M^2 R^4 G}, \frac{\varepsilon \cdot GM}{2MR^2}\right\}\right]$$

$$\leq 2\exp\left[-\frac{mn\varepsilon^2}{8R^4}\right]$$

where $m > 0$ is a universal constant and in the last step we used $G \geq n/2$ and w.l.o.g. we assumed that $\varepsilon \leq 2R^2$. Taking a union bound over all 9^d elements of $\mathcal{N}_{1/4}$, we get

$$\mathbb{P}\left[\left\|X_G S_G X_G^\top - G\mu \cdot I\right\|_2 \geq \varepsilon \cdot GM\right] \leq \mathbb{P}\left[\max_{\mathbf{v} \in \mathcal{N}_{1/4}} \left|\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v} - G\mu\right| \geq \frac{\varepsilon}{2} \cdot GM\right]$$

$$\leq 2 \cdot 9^d \exp\left[-\frac{mn\varepsilon^2}{8R^4}\right]$$

Setting $\varepsilon = 0.005c$ and noticing that $\mu \in [cM, M]$ by Lemma 13 finishes the proof. $\qquad\square$

Lemma 15 (Approximation Bound). *Consider two models $\mathbf{w}^1, \mathbf{w}^2 \in \mathbb{R}^d$ such that $\left\|\mathbf{w}^1 - \mathbf{w}^2\right\|_2 \leq \tau$ and let $\mathbf{s}^1, \mathbf{s}^2$ denote the M-truncated weight vectors they induce i.e. $s_i^j = \min\left\{M, \frac{1}{|\langle \mathbf{w}^j, \mathbf{x}_i\rangle - y_i|}\right\}, j = 1, 2$. Also let $S^1 = diag(\mathbf{s}^1)$ and $S^2 = diag(\mathbf{s}^2)$. Then for any $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ such that $\|\mathbf{x}_i\|_2 \leq R_X$ for all i,*

$$\left|\lambda_{\min}(XS^1 X^\top) - \lambda_{\min}(XS^2 X^\top)\right| \leq 2n\tau M^2 R_X^3$$

Proof. We have the following four cases with respect to the weights $s_i^j = \min\left\{M, \frac{1}{|\langle \mathbf{w}^j, \mathbf{x}_i\rangle - y_i|}\right\}, j = 1, 2$ these two models generate on any data point $\mathbf{x}_i \in \mathcal{B}_2(R_X)$. Note that we do not assume that these data points are generated from $\mathcal{D}$, just that they are bounded inside the ball $\mathcal{B}_2(R_X)$. Also note that although $\left|s_i^1 - s_i^2\right| \leq M$ trivially holds by virtue of truncation, such a result is not sufficient for us since our later analyses would like to be able to show $\left|s_i^1 - s_i^2\right| \leq \frac{M}{1000}$ by setting τ to be really small.

Case 1 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$ and $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| \leq \frac{1}{M}$. Here $s_i^1 = s_i^2 = M$ i.e. $\left|s_i^1 - s_i^2\right| = 0$.

Case 2 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$ but $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| \le \frac{1}{M}$. In this case $s_i^2 = M > s_i^1$. Thus,

$$\left|s_i^1 - s_i^2\right| = M - \frac{1}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i|} \le M - \frac{1}{|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i| + \tau R_X} \le M - \frac{M}{1 + \tau M R_X} < 2\tau M^2 R_X$$

Case 3 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| \le \frac{1}{M}$ but $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$. This is similar to Case 2 above.

Case 4 : $\left|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$ and $\left|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i\right| > \frac{1}{M}$. In this case we have

$$\left|\frac{1}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i|} - \frac{1}{|\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i|}\right| \le \frac{\left|\langle \mathbf{w}^1 - \mathbf{w}^2, \mathbf{x}_i\rangle\right|}{|\langle \mathbf{w}^1, \mathbf{x}_i\rangle - y_i| \cdot |\langle \mathbf{w}^2, \mathbf{x}_i\rangle - y_i|} \le 2\tau M^2 R_X$$

This tells us that $\left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \le 2n\tau M^2 R_X$. Now, if we let $S^1 = \mathrm{diag}(\mathbf{s}^1)$ and $S^2 = \mathrm{diag}(\mathbf{s}^2)$, then for any unit vector $\mathbf{v} \in S^{d-1}$, denoting $R_X := \max_{i\in[n]} \|\mathbf{x}_i\|_2$ we have

$$\left|\mathbf{v}^\top X S^1 X^\top \mathbf{v} - \mathbf{v}^\top X S^2 X^\top \mathbf{v}\right| = \left|\sum_{i=1}^n \left(\mathbf{s}_i^1 - \mathbf{s}_i^2\right) \langle \mathbf{x}_i, \mathbf{v}\rangle^2\right| \le \left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \cdot \max_{i\in[n]} \langle \mathbf{x}_i, \mathbf{v}\rangle^2 \le \left\|\mathbf{s}^1 - \mathbf{s}^2\right\|_1 \cdot R_X^2 \le 2n\tau M^2 R_X^3.$$

This proves that $\left\|X S^1 X^\top - X S^2 X^\top\right\|_2 \le 2n\tau M^2 R_X^3$ and concludes the proof. $\qquad\square$

D.1   Calculation of Distribution-specific Constants

The WSC/WSS bounds from Lemma 12 are parametrized by a constant c that lower bounds on the singular values of the matrix $X_G S_G X_G^\top$. Recall that for any covariate distribution $\mathcal{D}$, the constant is defined as

$$c := \inf_{\mathbf{u},\mathbf{v}\in S^{d-1}} \left\{ \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\min\left\{\frac{1}{|\langle \mathbf{u}, \mathbf{x}\rangle|}, 1\right\} \cdot \langle \mathbf{x}, \mathbf{v}\rangle^2\right]\right\}.$$

Below we present some interesting cases where this constant is lower bounded.

Centered Isotropic Gaussian For the special case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, I_d)$, notice that by rotational symmetry, we can, without loss of generality, take $\mathbf{u} = (1, 0, 0, \ldots, 0)$ and $\mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ where $v_1^2 + v_2^2 = 1$. Thus, if we consider $x_1, x_2 \sim \mathcal{N}(0, 1)$ i.i.d. then $c \ge \inf_{(v_1, v_2)\in S^1} f(v_1, v_2)$ where

$$\begin{aligned}
f(v_1, v_2) &= \mathbb{E}_{x_1, x_2 \sim \mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2 + 2v_1 v_2 x_1 x_2)\right]\\
&= \mathbb{E}_{x_1, x_2 \sim \mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2)\right]\\
&= \mathbb{E}_{x_1 \sim \mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2)\right]\\
&= \sqrt{\frac{2}{\pi}}\left(\int_0^1 (v_1^2 t^2 + v_2^2) e^{-t^2/2} dt + \int_1^\infty \left(v_1^2 t + \frac{v_2^2}{t}\right) e^{-t^2/2} dt\right)\\
&\ge 0.6827 \cdot v_1^2 + 0.9060 \cdot v_2^2
\end{aligned}$$

where in the second step we used the independence of x_1, x_2 and $\mathbb{E}[x_2] = 0$, in the third step we used independence once more and $\mathbb{E}\left[x_2^2\right] = 1$, and in the last step we used standard bounds on the error function and the exponential integral. This gives us $c \ge \inf_{(v_1, v_2)\in S^1} \left\{0.6827 \cdot v_1^2 + 0.9060 \cdot v_2^2\right\} \ge 0.68$.

Centered Non-isotropic Gaussian For the case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, \Sigma)$, we have $\mathbf{x} \sim \mathcal{D} \equiv \Sigma^{1/2} \cdot \mathcal{N}(\mathbf{0}, I_d)$. Thus, for any fixed unit vector $\mathbf{v}$, we have $\langle \mathbf{v}, \mathbf{x}\rangle \sim \langle \tilde{\mathbf{v}}, \mathbf{z}\rangle$ where $\tilde{\mathbf{v}} = \Sigma^{-1/2}\mathbf{v}$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I)$. We also have $\|\tilde{\mathbf{v}}\|_2 \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$ where $\lambda = \lambda_{\min}(\Sigma)$ and $\Lambda = \lambda_{\max}(\Sigma)$. Note that we must insist on

23

having $\lambda = \lambda_{\min}(\Sigma) > 0$ failing which, as the calculations show below, there is no hope of expecting c to be bounded away from 0. Now for any fixed vectors $\mathbf{u}, \mathbf{v}$ we first perform rotations so that we have $\tilde{\mathbf{u}} = (u, 0, 0, \ldots, 0)$ and $\tilde{\mathbf{v}} = (v_1, v_2, 0, 0, \ldots, 0)$ where we can assume w.l.o.g. that $u \geq 0$. Note that since $\{\|\tilde{\mathbf{u}}\|_2, \|\tilde{\mathbf{v}}\|_2\} \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$, we have $(v_1, v_2) \in S^r$ and $r, u \in \left[\frac{1}{\sqrt{\Lambda}}, \frac{1}{\sqrt{\lambda}}\right]$. This gives us $c \geq \inf_{(v_1,v_2) \in S^r} f(v_1, v_2)$ where

$$
\begin{aligned}
f(v_1, v_2) &= \mathop{\mathbb{E}}_{x_1,x_2 \sim \mathcal{N}(0,1)} \left[\min\left\{\frac{1}{u \cdot |x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2 + 2v_1 v_2 x_1 x_2)\right] \\
&= \mathop{\mathbb{E}}_{x_1,x_2 \sim \mathcal{N}(0,1)} \left[\min\left\{\frac{1}{u \cdot |x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2)\right] \\
&= \mathop{\mathbb{E}}_{x_1 \sim \mathcal{N}(0,1)} \left[\min\left\{\frac{1}{u \cdot |x_1|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2)\right] \\
&= \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(\int_0^{\frac{1}{u}} u(v_1^2 t^2 + v_2^2) e^{-t^2/2} dt + \int_{\frac{1}{u}}^{\infty} \left(v_1^2 t + \frac{v_2^2}{t}\right) e^{-t^2/2} dt\right) \\
&\geq \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(\int_0^{\frac{1}{u}} u(v_1^2 t^2 + v_2^2) e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} dt + v_1^2 e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} + \frac{v_2^2}{2} \int_{\frac{1}{2}\left(\frac{1}{u}\right)^2}^{\infty} \frac{1}{z} e^{-z} dz\right) \\
&\geq \frac{1}{u}\sqrt{\frac{2}{\pi}} \left(e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} \left(\frac{v_1^2}{3u^2} + v_2^2\right) + v_1^2 e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} + \frac{v_2^2}{4} e^{-\frac{1}{2}\left(\frac{1}{u}\right)^2} \log\left(1 + 4u^2\right)\right) \\
&\geq \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} \left(v_1^2\left(1 + \frac{\lambda}{3}\right) + v_2^2\left(1 + \frac{1}{4}\log\left(1 + \frac{4}{\Lambda}\right)\right)\right) \\
&\geq \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} (v_1^2 + v_2^2) \\
&= r^2 \sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}} \\
&\geq \frac{1}{\Lambda}\sqrt{\frac{2\lambda}{\pi}} e^{-\frac{\Lambda}{2}}
\end{aligned}
$$

where in the second and third steps we used independence of x_1, x_2, $\mathbb{E}[x_2] = 0$ and $\mathbb{E}[x_2^2] = 1$ as before, and in the sixth step we used lower bounds on the exponential integral.

Non-centered Isotropic Gaussian We discuss two techniques to handle the case of non-centered covariates.

- **Pairing Trick** This technique requires changes to the data points and relies on the fact that the difference of two i.i.d. non-centered Gaussian random variables is a centered Gaussian random variable with double the variance. Thus, given n covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, I_d)$ and corresponding responses $y_1, \ldots, y_n$, create $n/2$ data points (assume without loss of generality that n is even) $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \mathbf{x}_{i+n/2}}{\sqrt{2}}$ and $\tilde{y}_i = \frac{y_i - y_{i+n/2}}{\sqrt{2}}$. Clearly $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, 2 \cdot I_d)$. However, this method has drawbacks since it is likely to increase the proportion of corrupted data points. If α fraction of the original points were corrupted, at most 2α fraction of the new points would be corrupted.

- **Direct Centering** Suppose we have data from a distribution $\mathcal{D} = \mathcal{N}(\boldsymbol{\mu}, I_d)$. As earlier, by rotational symmetry, we can take $\mathbf{u} = (1, 0, 0, \ldots, 0)$, $\mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, 0, 0, \ldots, 0)$. Assume $\|\boldsymbol{\mu}\|_2 = \rho$ and, without loss of generality, $\rho \geq 2$. Letting $\langle \boldsymbol{\mu}, \mathbf{v} \rangle =: p \leq \rho$ and $x_1, x_2, x_3 \sim \mathcal{N}(0, 1)$ i.i.d. gives $c \geq \inf_{(v_1,v_2) \in S^1} f(v_1, v_2)$ where, as before, independence of x_1, x_2, x_3 and the fact that $\mathbb{E}[x_2] = 0$ and $\mathbb{E}[x_2^2] = 1$, gives us

$$
f(v_1, v_2) = \mathop{\mathbb{E}}_{x_1 \sim \mathcal{N}(0,1)} \left[\min\left\{\frac{1}{|x_1 + \mu_1|}, 1\right\} \cdot ((p + v_1 x_1)^2 + v_2^2)\right]
$$

24

Now, since $(v_1, v_2) \in S^1$ we get two cases (recall that we have assumed w.l.o.g. $\rho \geq 2$)

Case 1: $v_2^2 \geq \frac{1}{2}$ In this case $f(v_1, v_2) \geq \frac{1}{2} \mathbb{E}_{x_1 \sim \mathcal{N}(0,1)} \left[\min \left\{ \frac{1}{|x_1 + \mu_1|}, 1 \right\} \right] \geq \Omega \left(\exp^{-\rho^2/2} \log \left(1 + \frac{1}{\rho^2} \right) \right)$.

Case 2: $v_1^2 \geq \frac{1}{2}$ In this case, if $x_1 \geq 2\sqrt{2}\rho$, then $|v_1 x_1 + p| \geq \frac{v_1 x_1}{2}$, as well as $|x_1 + \mu_1| \leq 2x_1$.

$$
\begin{aligned}
f(v_1, v_2) &\geq \mathbb{E}_{x_1 \sim \mathcal{N}(0,1)} \left[\min \left\{ \frac{1}{|x_1 + \mu_1|}, 1 \right\} (p + v_1 x_1)^2 \cdot \mathbb{I} \left\{ x_1 \geq 2\sqrt{2}\rho \right\} \right] \\
&\geq \mathbb{E}_{x_1 \sim \mathcal{N}(0,1)} \left[\min \left\{ \frac{1}{2x_1}, 1 \right\} \frac{x_1^2}{8} \cdot \mathbb{I} \left\{ x_1 \geq \max 2\sqrt{2}\rho \right\} \right] \geq \frac{1}{16} e^{-4\rho^2}
\end{aligned}
$$

Since the value ρ influences the final bound on c very heavily, it is advisable to avoid a large ρ value. One way to ensure this is to algorithmically center the covariates i.e. use $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \hat{\boldsymbol{\mu}}$ where $\hat{\boldsymbol{\mu}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. This would (approximately) center the covariates and ensure an effective value of $\rho \approx \mathcal{O} \left(\sqrt{\frac{d}{n}} \right)$

Bounded Sub-Gaussian Suppose our covariate distribution has bounded support i.e. $\text{supp}(\mathcal{D}) \subset \mathcal{B}_2(\rho)$ for some $\rho > 0$. Assume $\rho \geq 1$ w.l.o.g. Also, using the centering trick above, assume that $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{x}] = \mathbf{0}$. Then we have $|\langle \mathbf{u}, \mathbf{x} \rangle| \leq \rho$ which implies $\min \left\{ \frac{1}{|\langle \mathbf{u}, \mathbf{x} \rangle|}, 1 \right\} \geq \frac{1}{\rho}$. Let Σ denote the covariance of the distribution $\mathcal{D}$ and let $\lambda := \lambda_{\min}(\Sigma)$ denote its smallest eigenvalue. This gives us $c \geq \frac{1}{\rho} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\langle \mathbf{x}, \mathbf{v} \rangle^2 \right] \geq \frac{\lambda}{\rho}$.

E   Corruptions and Dense Noise - Supplementary Details

In this section, we will provide details of the convergence analysis of STIR and STIR-GD in the setting where even the "good" points experience sub-Gaussian noise. Thus, we will assume that our data is generated as $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b} + \boldsymbol{\epsilon}$ where, as before $\|\mathbf{b}\|_0 \leq \alpha \cdot n$ and $\boldsymbol{\epsilon} \sim \mathcal{D}_\varepsilon$ where $\mathcal{D}_\varepsilon$ is a σ-sub-Gaussian distribution with zero mean and real support. As mentioned before, we can tolerate noise with non-zero mean as well, by using the same pairing trick we used to center the covariates in Appendix D.1. This would have a side effect of at most doubling the corruption rate α. We will denote, as before $B := \text{supp}(\mathbf{b})$ and $G := [n] \setminus B$. Our covariates will continue to be sampled from an R sub-Gaussian distribution $\mathcal{D}$ with support over $\mathbb{R}^d$. We (re)state the main result of this section below.

Theorem 2. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted with the rest subjected to sub-Gaussian noise sampled from a distribution $\mathcal{D}_\varepsilon$ with sub-Gaussian norm σ. If STIR (or STIR-GD) is initialized at an (arbitrary) point $\mathbf{w}^0$, with an initial truncation that satisfies $M_1 \leq \frac{1}{\|\mathbf{w}^0 - \mathbf{w}^*\|_2}$, and executed with an increment $\eta > 1$ such that we have $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$, where $c_\varepsilon > 0$ is a constant that depends only on the distributions $\mathcal{D}$ and $\mathcal{D}_\varepsilon$, then with probability at least $1 - \exp \left(-\Omega \left(n - d \log(d + n) + \log \frac{1}{M_1 \sigma} \right) \right)$, after $K = \mathcal{O} \left(\log \frac{1}{M_1 \sigma} \right)$ stages, each of which has only $\mathcal{O}(1)$ iterations, we must have $\|\mathbf{w}^K - \mathbf{w}^*\|_2 \leq \mathcal{O}(\sigma)$.*

Proof. The overall proof of this result follows exactly the same way as the result in Theorem 1. We will still utilize the notion of a *well-initialized stage* and establish (see Lemma 16 below) a convergence guarantee for each well-initialized stage. However, Lemma 16 will itself require a few new results to be proved.

However, note that Lemma 8, a similar result for well-initialized stages in the setting without dense noise, required two results, namely Lemmata 12 and 10 that established the WSC/WSS properties and bounded the weight put on bad points. Those results implicitly assumed that good points incur absolutely no modification to their response value which is no longer true here since in the setting being considered here, even good points do incur sub-Gaussian noise in their responses. Thus, we will establish below Lemmata 17 and 18 which will establish those results in the dense noise setting. We note that a similar convergence guarantee may be established for STIR-GD in the dense noise setting as well.

However, note that this result only guarantees a convergence to $\left\|\mathbf{w}^{K,1} - \mathbf{w}^*\right\|_2 \leq \mathcal{O}\left(\sigma\right)$ and thus, does not ensure a consistent solution. A technical reason for this is because Lemma 17 holds true only for values of $M \leq \mathcal{O}\left(\frac{1}{\sigma}\right)$ which restricts the application of this result to offer errors much smaller than σ. It would be interesting to show, as [5] do, that STIR, or a variant, does offer consistent estimates.

For sake of notational simplicity, we will assume that $\boldsymbol{\epsilon}_B = \mathbf{0}$ by shifting any sub-Gaussian noise a bad point, say $j \in B$ does incur, into the corruption value corresponding to that point i.e. $\mathbf{b}_j$. This is without loss of generality since we impose no constraints on the corruptions other than that they be sparse, in particular the corruptions need not be bounded and can thus, absorb sub-Gaussian noise values into them. $\qquad\square$

Lemma 16. *Suppose we have n data points with the covariates $\mathbf{x}_i$ sampled from a sub-Gaussian distribution $\mathcal{D}$ and an α fraction of the data points are corrupted with the rest experiencing noise generated i.i.d. from a distribution $\mathcal{D}_\varepsilon$ with sub-Gaussian norm σ. Suppose we initialize a stage T within an execution of STIR with truncation level $M \leq \frac{c_\varepsilon}{8\eta\sigma}$, increment parameter η, and a model $\mathbf{w}^T =: \mathbf{w}^T T, 1$ such that $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$ and $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{1}{M}$, then with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, there exists an upper bound of $t_0 = \mathcal{O}\left(1\right)$ iterations, such that we are assured that $\left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2 \leq \frac{1}{\eta M}$ for all $\tau \geq t_0$. Here c_ε is the constant of the WSC property and depends only on the distributions $\mathcal{D}$ and $\mathcal{D}_\varepsilon$ (see Lemma 17).*

Proof. Let $\mathbf{w}^{T,\tau}$ be a model encountered by STIR within this stage and let $\mathbf{r} = X^\top \mathbf{w}^{T,\tau} - \mathbf{y}$ denote the residuals due to $\mathbf{w}^{T,\tau}$ and $S = \text{diag}(\mathbf{s})$ denote the diagonal matrix of weights where $\mathbf{s}_i = \min\left\{\frac{1}{|\mathbf{r}_i|}, M\right\}$. Then STIR will choose as the next model $\mathbf{w}^{T,\tau+1} = (XSX^\top)^{-1}XS\mathbf{y} = \mathbf{w}^* + (XSX^\top)^{-1}XS(\mathbf{b} + \boldsymbol{\epsilon})$ which gives us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{\|XS(\mathbf{b} + \boldsymbol{\epsilon})\|_2}{\lambda_{\min}(XSX^\top)}$$

Now by Lemma 5, with probability at least $1 - \exp(-\Omega\left(n - d\right))$, we have $\|X_B\|_2 = \sqrt{\lambda_{\max}(X_B X_B^\top)} \leq \sqrt{1.01B}$. By Lemma 10, with the same probability, we have

$$\|S\mathbf{b}\|_2 \leq \sqrt{4B(1 + 1.01M^2 \|\mathbf{w} - \mathbf{w}^*\|_2^2)} \leq 2\sqrt{2.01B},$$

whereas by Lemma 18, as we have restricted $M \leq \frac{1}{8\sigma}$, we have, yet again with the same probability,

$$\|XS\boldsymbol{\epsilon}\|_2 = \|X_G S_G \boldsymbol{\epsilon}_G\| \leq 4MG\sigma\sqrt{1.01} \leq \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G,$$

where the first equality follows due to our convention that $\text{supp}(\boldsymbol{\epsilon}) = G$ since for bad points in the set B, we clubbed any sub-Gaussian noise into the corruption itself, thus leaving $\boldsymbol{\epsilon}_B = \mathbf{0}$. Now, by Lemma 17, with probability at least $1 - \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$, we have $\lambda_{\min}(XSX^\top) \geq \lambda_{\min}(X_G S_G X_G^\top) \geq 0.99c_\varepsilon \cdot GM$. This give us

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq \frac{2B\sqrt{2.0301} + \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G}{0.99c_\varepsilon \cdot GM} \leq \frac{2B\sqrt{2.0301}}{0.99c_\varepsilon \cdot GM} + \frac{\sqrt{1.01}}{1.98\eta \cdot M}$$

Now, since we have $\alpha \leq \frac{c_\varepsilon}{5.85\eta + c_\varepsilon}$, we also have $\frac{2B\sqrt{2.0301}}{0.99c_\varepsilon \cdot GM} \leq \left(1 - \frac{\sqrt{1.01}}{1.98}\right)\frac{1}{\eta M}$ and thus, $\frac{2B\sqrt{2.0301} + \frac{c_\varepsilon\sqrt{1.01}}{2\eta}G}{0.99c_\varepsilon \cdot GM} \leq \frac{1}{\eta M}$. Arguing as we did in the proof of Lemma 8, we must either have $\|\mathbf{w}^+ - \mathbf{w}^*\|_2 \leq \frac{2B\sqrt{2.0301}}{0.9801c_\varepsilon \cdot GM} + \frac{\sqrt{1.01}}{1.9602\eta \cdot M}$ and if that does not happen, we must instead have

$$\left\|\mathbf{w}^{T,\tau+1} - \mathbf{w}^*\right\|_2 \leq 0.99 \cdot \left\|\mathbf{w}^{T,\tau} - \mathbf{w}^*\right\|_2$$

This proves the claimed result. $\qquad\square$

E.1 Establishing WSC/WSS in Presence of Dense Noise

We will rework a counterpart to Lemma 12 in this section.

Lemma 17. *Given the problem setting above, then there exists a constant $c_\varepsilon > 0$ that depends only on the distributions $\mathcal{D}, \mathcal{D}_\varepsilon$ such that for any $M \in \left[0, \frac{1}{\sigma}\right]$, we have*

$$\mathbb{P}\left[\exists S \in \mathcal{S}_M\left(\frac{1}{M}\right) : \lambda_{\min}(X_G S_G X_G^\top) < 0.99 c_\varepsilon \cdot GM\right] \leq \exp\left(-\Omega\left(n - d\log(d+n)\right)\right)$$

In particular, for standard Gaussian covariates and Gaussian noise with variance σ^2, we can take $c_\varepsilon \geq 0.52$.

Proof. Let $\mathbf{x} \sim \mathcal{D}, \epsilon \sim \mathcal{D}_\varepsilon$ and let $y = \langle \mathbf{w}^*, \mathbf{x}\rangle + \epsilon$ be the response of an uncorrupted data point and $\mathbf{w} \in \mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ be any fixed model. Then if we let $\boldsymbol{\Delta} := \mathbf{w} - \mathbf{w}^*$, the weight s that the model $\mathbf{w}$ would cause STIR to put on this (clean) data point must satisfy $s \geq \min\left\{\frac{1}{|\langle \boldsymbol{\Delta}, \mathbf{x}\rangle - \epsilon|}, M\right\}$. This gives us, for any fixed $\mathbf{v} \in S^{d-1}$,

$$\mathbb{E}\left[\mathbf{v}^\top X_G S_G X_G^\top \mathbf{v}\right] \geq c_\varepsilon \cdot GM,$$

where we define,

$$c_\varepsilon := \inf_{\substack{0 \leq r \leq \frac{1}{M} \\ \mathbf{u}, \mathbf{v} \in S^{d-1}}} \left\{ \mathbb{E}_{\mathbf{x}\sim\mathcal{D}, \epsilon\sim\mathcal{D}_\varepsilon}\left[\min\left\{\frac{1}{|Mr\langle\mathbf{u}, \mathbf{x}\rangle - M\epsilon|}, 1\right\} \cdot \langle\mathbf{x}, \mathbf{v}\rangle^2\right]\right\}$$

We analyze the constant c for the Gaussian case at the end of the proof. For now, we proceed as in Lemma 14 and realize that the sub-Gaussian norm calculations continue to hold in this case since they simply upper bound the weights by M, and get

$$\mathbb{P}\left[\lambda_{\min}(X_G S_G X_G^\top) < 0.995 c_\varepsilon \cdot GM\right] \leq 2 \cdot 9^d \exp\left[-\frac{mn(0.005 c_\varepsilon)^2}{8R^4}\right]$$

After this we notice that the proof of Lemma 15 pays no heed to corruptions or additional noise and hence, continues to hold in this setting too. Proceeding as in the proof of Lemma 12 to set up a τ-net over $\mathcal{B}_2\left(\mathbf{w}^*, \frac{1}{M}\right)$ and taking a union bound over this net finishes the proof.

For the special case of $\mathcal{D} = \mathcal{N}(\mathbf{0}, I_d)$ and $\mathcal{D}_\varepsilon = \mathcal{N}(0, \sigma^2)$, by rotational symmetry, we can, without loss of generality, take $\mathbf{u} = (1, 0, 0, \ldots, 0)$ and $\mathbf{v} = (v_1, v_2, 0, 0, \ldots, 0)$ where $v_1^2 + v_2^2 = 1$. Thus, if $x_1, x_2, \epsilon \sim \mathcal{N}(0,1)$ i.i.d. then $c \geq \inf_{(v_1, v_2)\in S^1, r\in\left[0, \frac{1}{M}\right]} f(v_1, v_2, r)$ where

$$
\begin{aligned}
f(v_1, v_2, r) &= \mathbb{E}_{x_1, x_2, \epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2 + 2v_1 v_2 x_1 x_2)\right] \\
&= \mathbb{E}_{x_1, x_2, \epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2 x_2^2)\right] \\
&= \mathbb{E}_{x_1, \epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} \cdot (v_1^2 x_1^2 + v_2^2)\right] \\
&= v_1^2 \cdot \underbrace{\mathbb{E}_{x_1, \epsilon\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{|Mrx_1 - M\sigma\epsilon|}, 1\right\} x_1^2\right]}_{(A)} + v_2^2 \cdot \underbrace{\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{M\sqrt{r^2 + \sigma^2}|z|}, 1\right\}\right]}_{(B)}
\end{aligned}
$$

where in the second step we used the independence of x_1, x_2 and $\mathbb{E}[x_2] = 0$, in the third step we used independence once more and $\mathbb{E}\left[x_2^2\right] = 1$. In the fourth step, we substituted $\sqrt{r^2 + \sigma^2} z = rx_1 - \sigma\epsilon$ and noticed that $rx_1 - \sigma\epsilon \sim \mathcal{N}(0, (r^2 + \sigma^2))$ i.e. $z \sim \mathcal{N}(0,1)$. To bound (B) we notice $r \leq \frac{1}{M}$ and $M \leq \frac{1}{\sigma}$ and use standard bounds on Gaussian and exponential integrals to get

$$(B) \geq \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\min\left\{\frac{1}{\sqrt{2}|z|}, 1\right\}\right] \geq 0.815$$

27

To bound (A), we use the fact that pairwise orthogonal projections of a standard Gaussian vector yield independent variables. Thus, if we denote $a = Mr, b = M\sigma$ and $z = \frac{ax_1 - b\epsilon}{\sqrt{a^2 + b^2}}, w = \frac{bx_1 + a\epsilon}{\sqrt{a^2 + b^2}}$, then $z, w \sim \mathcal{N}(0, 1)$ as well as $z \perp w$. Thus, we have

$$
\begin{aligned}
(A) &= \mathop{\mathbb{E}}_{z, w \sim \mathcal{N}(0,1)} \left[\min \left\{ \frac{1}{M\sqrt{r^2 + \sigma^2}\,|z|}, 1 \right\} \cdot \left(\frac{r^2 z^2 + \sigma^2 w^2 + 2r\sigma zw}{r^2 + \sigma^2} \right) \right] \\
&\geq \mathop{\mathbb{E}}_{z, w \sim \mathcal{N}(0,1)} \left[\min \left\{ \frac{1}{\sqrt{2}\,|z|}, 1 \right\} \cdot \left(\frac{r^2 z^2 + \sigma^2 w^2}{r^2 + \sigma^2} \right) \right] \\
&\geq \frac{0.52 r^2}{r^2 + \sigma^2} + \frac{0.815 \sigma^2}{r^2 + \sigma^2} = 0.52 + \frac{0.295 \sigma^2}{r^2 + \sigma^2}
\end{aligned}
$$

where in the second step we used $M \leq \frac{1}{\sigma}$ and $r \leq \frac{1}{M}$, independence of z and w and the fact that $\mathbb{E}[w] = 0, \mathbb{E}[w^2] = 1$ and the last step uses standard bounds on Gaussian and exponential integrals. $\qquad\square$

E.2 Bounding the Weights on Good Points

Although Lemma 10 continues to hold in this case, since good points also incur modifications to their response values, albeit modifications that are stochastic and not adversarial, we need an analogous result for the good points in this case as well.

Lemma 18. *Suppose σ is the sub-Gaussian norm of the noise distribution $\mathcal{D}_\varepsilon$ and the identity of the good points G is chosen independently of the covariates. Then for any $M > 0$, if S is the diagonal matrix of M-truncated weights assigned to the data points by a model $\mathbf{w}$, then with probability at least $1 - \exp(-\Omega(n - d))$,*

$$
\|X_G S_G \boldsymbol{\epsilon}_G\|_2 \leq 4MG\sigma\sqrt{1.01}
$$

Proof. We have, by applying Lemma 5, with probability at least $1 - \exp(-\Omega(n - d))$,

$$
\|X_G S_G \boldsymbol{\epsilon}_G\|_2 \leq \sqrt{\lambda_{\max}(X_G X_G^\top)} \cdot \|S_G \boldsymbol{\epsilon}_G\| \leq \sqrt{1.01 G} \cdot \|S\|_2 \|\boldsymbol{\epsilon}_G\|_2 \leq \sqrt{1.01 G} M \cdot \|\boldsymbol{\epsilon}_G\|_2 ,
$$

where the last inequality follows since S is a diagonal matrix and by M-truncation, the maximum value of any weight is M. Now, since our noise is σ sub-Gaussian and unbiased, we have, for any fixed $\mathbf{u} \in S^{G-1}$, $\mathbb{E}[\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle] = 0$, as well as, by applying the Hoeffding's inequality,

$$
\mathbb{P}[|\langle \boldsymbol{\epsilon}, \mathbf{u} \rangle| \geq t] \leq 3 \exp\left(-\frac{t^2}{2\sigma^2} \right)
$$

Now, if $\mathbf{u}^1, \mathbf{u}^2 \in S^{G-1}$, such that $\|\mathbf{u}^1 - \mathbf{u}^2\|_2 \leq \frac{1}{2}$, then we have $|\langle \mathbf{u}^1 - \mathbf{u}^2, \boldsymbol{\epsilon} \rangle| \leq \frac{1}{2} \cdot \|\boldsymbol{\epsilon}\|_2$. Thus, taking a union bound over a $1/2$-net over S^{G-1} gives us

$$
\mathbb{P}\left[\|\boldsymbol{\epsilon}\|_2 = \max_{\mathbf{u} \in S^{G-1}} \langle \mathbf{u}, \boldsymbol{\epsilon} \rangle \geq \frac{1}{2} \cdot \|\boldsymbol{\epsilon}\|_2 + t \right] = \mathbb{P}[\|\boldsymbol{\epsilon}\|_2 \geq 2t] \leq 3 \cdot 5^G \exp\left[-t^2/2\sigma^2 \right]
$$

Setting $t = \sigma\sqrt{4G}$ establishes the result. $\qquad\square$

F Robust Linear Bandits

In this section, we briefly discuss the linear contextual bandit problem with corrupted arm pulls. We refer the reader to [19] for a more relaxed introduction to the problem as well as formal regret bounds. Indeed, the discussion here is adapted from the discussion in [19].

Problem Setting 1 Adversarial Linear Bandits

---

for $t = 1, 2, 3..$ **do**
    Player receives a set of contexts $A_t = \left\{\mathbf{x}^{t,1}, \ldots, \mathbf{x}^{t,n_t}\right\} \subset \mathbb{R}^d$
    Player plays an arm, $\hat{\mathbf{x}}^t \in A_t$
    Clean reward is generated $r_t^* = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$ conditioned on $\mathcal{H}^t$
    Adversary inspects $\hat{\mathbf{x}}^t, r_t^*, \mathcal{H}^t$ and chooses b_t //`while making sure` $|\tau \leq t : b_\tau \neq 0| \leq \eta \cdot (t+1)$
    Player receives reward, $r_t = r_t^* + b_t$
end for

F.1   Problem Setting

The stochastic linear contextual bandit framework [1, 20] considers a (possibly infinite) set of *arms*. Arms correspond to various actions that can be performed by the algorithm. For instance, in a recommendation setting, arms may correspond to various products that are available for sale, for instance, at an e-commerce website, or in a quantitative trading setting, arms may correspond to stocks that are available for sale/purchase.

Every arm $\mathbf{a}$ is parametrized by a vector $\mathbf{a} \in \mathbb{R}^d$ (we abuse notation to denote the arm and its corresponding parametrization using the same notation). Recall that the set of all arms is potentially infinite. However, not all arms may be available at every time step. For instance, an e-commerce website would not like to recommend products not currently in stock. Similarly, stocks not currently in one's possession cannot be sold.

At each time step t, the algorithm receives a set of n_t arms (also called *contexts*) $A_t = \left\{\mathbf{x}^{t,1}, \ldots, \mathbf{x}^{t,n_t}\right\} \subset \mathbb{R}^d$ that can be played or *pulled* in this round. Pulling an arm is akin o performing the action associated with that arm, for example, recommending an item or selling a stock unit. The context set A_t, as well as the number n_t of contexts available can vary across time steps. The algorithm selects and pulls an arm $\hat{\mathbf{x}}^t \in A_t$ as per its arm selection policy. In response, a reward r_t is generated. Let $\mathcal{H}^t = \left\{A_1, \hat{\mathbf{x}}^1, r_1, \ldots, A_{t-1}, \hat{\mathbf{x}}^{t-1}, r_{t-1}, A_t, \hat{\mathbf{x}}^t\right\}$.

F.2   Adversary Model

In the stochastic linear bandit setting, as has been studied in prior work [1, 20] , at every time step, the reward r_t is generated using a *model vector* $\mathbf{w}^* \in \mathbb{R}^d$ (that is not known to the algorithm) as follows: $r_t = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$, where ϵ_t is a *noise* value that is typically assumed to be (conditionally) centered and σ-sub-Gaussian, i.e., $\mathbb{E}\left[\epsilon_t \mid \mathcal{H}^t\right] = 0$, as well as for some $\sigma > 0$, we have $\mathbb{E}\left[\exp(\lambda\epsilon_t) \mid \mathcal{H}^t\right] \leq \exp(\lambda^2\sigma^2/2)$ for any $\lambda > 0$.

However, recent works [19, 22] have considered settings where the rewards may suffer not only sub-Gaussian noise, but also adversarial corruptions that are introduced by an *adaptive adversary* that is able to view the on-goings of the online process and at any time instant t, *after* observing the history $\mathcal{H}^t$ and the "clean" reward value, i.e., $\langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t$, is able to add a corruption value b_t to the reward. For notational uniformity, we will assume that for time instants where the adversary chooses not to do anything, $b_t = 0$. Thus, the final reward to the player at every time step is $r_t = \langle \mathbf{w}^*, \hat{\mathbf{x}}^t \rangle + \epsilon_t + b_t$. This model is described in Problem Setting 1.

For sake of simplicity we will assume that, for some $B > 0$, the final (possibly corrupted) reward presented to the player satisfies $r_t \in [-B, B]$ almost surely. The only constraint the adversary need observe while introducing the corruptions is that at no point in the online process, should the adversary have corrupted more than an η fraction of the observed rewards. Formally, let $G_t = \{\tau < t : b_\tau = 0\}$ and $B_t = \{\tau < t : b_\tau \neq 0\}$ denote the set of "good" and "bad" time instances till time t. We insist that $|B_t| \leq \eta \cdot t$ for all t.

Algorithm 4 WUCB-Lin: Weighted UCB for Linear Contextual Bandits

---

Input: Upper bounds σ_0 (on sub-Gaussian norm of noise distribution), B (on magnitude of corruption), α_0 (on fraction of corrupted points), initial truncation M_1, increment rate η

1: **for** $t = 1, 2, \ldots, T$ **do**
2:     Receive set of arms A_t
3:     Play arm $\hat{\mathbf{x}}^t = \underset{\mathbf{x} \in A_t, \mathbf{w} \in C_{t-1}}{\arg \max} \langle \mathbf{x}, \mathbf{w} \rangle$
4:     Receive reward r_t
5:     $(\hat{\mathbf{w}}^t, S^t) \leftarrow \mathsf{STIR}\left(\{\hat{\mathbf{x}}^\tau, r_\tau\}_{\tau=1}^t, M_1, \eta\right)$ `//Denote` $S^t = \mathtt{diag}(s_1^t, s_2^t, \ldots, s_t^t)$
6:     $V^t \leftarrow \sum_{\tau \leq t} s_\tau^t \hat{\mathbf{x}}^\tau (\hat{\mathbf{x}}^\tau)^\top, \; X^t \leftarrow \left[\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2, \ldots, \hat{\mathbf{x}}^t\right]$
7:     $\bar{\mathbf{w}}^t \leftarrow (V^t)^{-1} X^t S^t \mathbf{y}$
8:     $C_t \leftarrow \{\mathbf{w} : \left\|\mathbf{w} - \bar{\mathbf{w}}^t\right\|_{V^t} \leq \sigma_0 \sqrt{d \log T} + \alpha_0 B T\}$
9: **end for**

F.3   Notion of Regret

The goal of the algorithm is to maximize the cumulative reward it receives over the time steps $\sum_{t=1}^T r_t$. However, a more popular technique of casting this objective is in the form of *cumulative pseudo regret*. At time t, let $\mathbf{x}^{t,*} = \arg\max_{\mathbf{x} \in A_t} \langle \mathbf{w}^*, \mathbf{x} \rangle$ be the arm among those available that yields the highest expected (uncorrupted) reward. The cumulative pseudo regret of a policy π is defined as follows

$$\bar{R}_T(\pi) = \sum_{t=1}^T \langle \mathbf{w}^*, \mathbf{x}^{t,*} \rangle - \mathbb{E}\left[r_t\right].$$

Note that the best arm here may change across time-steps.

F.4   WUCB-Lin: An Algorithm for Robust Linear Bandits

We use the notation $\|\mathbf{x}\|_M = \sqrt{\mathbf{x}^\top M \mathbf{x}}$ for a vector $\mathbf{x} \in \mathbb{R}^d$ and a matrix $M \in \mathbb{R}^{d \times d}$. We reproduce, for convenience, the WUCB-Lin algorithm in Algorithm 4. WUCB-Lin builds upon the OFUL principle [1] for linear contextual bandits. At every step, WUCB-Lin uses rewards obtained from previous arm pulls to obtain an estimate $\hat{\mathbf{w}}^t$ of the true model vector $\mathbf{w}^*$.

Whereas classical algorithms utilize ordinary least squares to solve this problem, WUCB-Lin utilizes STIR (actually STIR-GD for sake of speed) to obtain this estimate. This lends resilience to the algorithm against the (possibly several) past arm pulls whose rewards got corrupted by the adversary. The previous work of [19] used the TORRENT algorithm for the same purpose.

The next step in executing the OFUL principle is the construction of a *confidence set*. It is common to use an ellipsoidal confidence set with the ellipsoid induced by the covariance matrix of the arm vectors pulled so far. The work of [19] modifies this to only consider arms considered as clean by the TORRENT algorithm while constructing the confidence ellipsoid.

Since STIR, instead of selecting a specific subset of arms like TORRENT, instead would assign weights to all previously pulled arms, with a small weight indicating a high likelihood of the arm pull being a corrupted one and a large weight indicating a high likelihood of the arm pull being a clean one. Thus, STIR utilizes these weights to construct a *weighted covariance matrix* which is then used to define the confidence ellipsoid and carry out the arm selection step.