Efficient and Consistent Robust Time Series Analysis

Kush Bhatia* Prateek Jain* Parameswaran Kamalaruban#
Purushottam Kar†
*Microsoft Research, Bangalore, India
{t-kushb, prajain}@microsoft.com
#Australian National University, Canberra, Australia
kamalaruban.parameswaran@nicta.com.au
†Indian Institute of Technology Kanpur, India
purushot@cse.iitk.ac.in

Abstract

We study the problem of robust time series analysis under the standard auto-regressive (AR) time series model in the presence of arbitrary outliers. We devise an efficient hard thresholding based algorithm which can obtain a *consistent* estimate of the optimal AR model despite a large fraction of the time series points being corrupted. Our algorithm alternately estimates the corrupted set of points and the model parameters, and is inspired by recent advances in robust regression and hard-thresholding methods. However, a direct application of existing techniques is hindered by a critical difference in the time-series domain: each point is correlated with *all* previous points rendering existing tools inapplicable directly. We show how to overcome this hurdle using novel proof techniques. Using our techniques, we are also able to provide the *first* efficient and provably consistent estimator for the robust regression problem where a standard linear observation model with white additive noise is corrupted arbitrarily. We illustrate our methods on synthetic datasets and show that our methods indeed are able to consistently recover the optimal parameters despite a large fraction of points being corrupted.

1 Introduction

Several real world prediction problems, for instance, the temperature of a city, stock prices, traffic patterns, the GPS location of a car etc are naturally modeled as time series. One of the most popular and simple model for time series is the auto-regressive (AR (d)) model which models a given observation as a sample from a distribution with mean given by a fixed linear combination of previous d time series values. That is, $x_t = \sum_{i=1}^{d} w_i^* x_{t-i} + \epsilon_i$ where ϵ_i is unbiased noise.

Unfortunately, in real life scenarios, time series tend to have several outliers. For example, traffic patterns may get disrupted due to accidents and stock prices may get affected by unforseen political or social influences. The estimation of model parameters in the presence of such outliers is a classical problem in time-series literature and is given a detailed treatment in several texts [11, 12].

Existing time-series texts define two major outlier models: a) innovative outliers, b) additive outliers. In innovative outliers, corrupted values become a part of the time series and influence future iterates i.e. if x_t is corrupted and we observe $\tilde{x}_t = x_t + b_t$ then subsequent values $x_{t'}$ $(t' > t)$ are obtained by using $\tilde{x}_t$ rather than x_t. In the additive outlier model, on the other hand, although the observation of $\tilde{x}_t$ is corrupted, the time series itself continues using the clean value x_t. Conventional wisdom in time series literature considers innovative outliers to be "good" and helpful in spurring a shift in the time series [11]. Additive outliers, on the other hand, are considered more challenging due to this latent behaviour in the model and can cause standard estimators for the AR model to diverge.

Due to importance of the problem, several estimators have been proposed for the AR model under corruption, e.g. the generalized M-estimator by [13]. However, most existing estimators are computationally intractable (operate in exponential time) and do not offer non-asymptotic guarantees.

Our goal in this work is to devise an *efficient* and *consistent* estimator for the Robust Time Series Estimation (RTSE) problem in the AR(d) model with non-asymptotic convergence guarantees in the presence of a large number of outliers. To this end, we cast the model estimation problem as a sparse estimation problem and use techniques from the sparse regression literature [10] to devise our hard-thresholding based algorithm. At a high level, our algorithm locates the corrupted indices by using a projected gradient method where the projection is onto the set of sparse vectors.

However, analyzing this technique proves especially challenging. While hard threshodling methods have been extensively studied for sparse linear regression [10, 6, 23], similar techniques do not apply directly to our problem because of two key challenges: a) in the time series domain, data points x_t's are dependent on each other while sparse linear regression techniques typically assume independence of the data points, and b) even for robust linear regression (where each row of data matrix is assumed to be independent), existing analyses [5] are unable to guarantee consistent estimates.

Using a novel two-stage proof technique, we show that our method provides a consistent estimator for the true model $\mathbf{w}^*$ so long as the number of outliers k satisfies $k = O(\frac{n}{d \log n})$, where n is the total number of points in the time series and d is the order of the model. Whenever k satisfies the above assumption, our method in time $\tilde{O}(nd)$ outputs an estimate $\widehat{\mathbf{w}}$ s.t. $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq f(n)$ where $f \to 0$ as $n \to \infty$. We direct the reader to Theorem 9 for precise rates.

In fact, using our techniques, we are also able to give a *consistent* estimator for the robust least squares regression (RLSR) problem [22, 16, 5] even when a constant fraction of the responses are corrupted. Here again, our algorithm runs in time $\tilde{O}(nd)$, where d is the dimensionality of the data. To the best of our knowledge, our method is the *first* efficient and consistent estimator for the RLSR problem in the challenging setting where a *constant* fraction of the responses can be corrupted.

We then study our methods empirically for both the robust time series analysis, as well as the standard robust regression problems. Our methods demonstrate consistency for both problem settings. Moreover, our results for robust time series show that the ordinary least squares estimate, that ignores outliers, provides very poor estimators and hence, is significantly less accurate. In contrast, our proposed method and a few variants of it indeed recover the underlying AR(d) model accurately.

Paper Organization: Section 3 considers the "warm-up" problem of robust regression and presents our algorithm and theoretical guarantees. We then, introduce the robust time series problem and our algorithm and analysis in Section 4. Section 5 presents simulations on synthetic datasets.

2 Related Works

Time Series: Analysing time series with corruptions is a classical and widely studied problem in statistics literature. In an early work, [13] proposed a generalized M-estimator for the RTSE problem in the additive outlier (AO) model with a positive breakdown point. [11] detail a robust variant of the Durbin-Levinson algorithm for RTSE and demonstrate the efficacy of the model empirically. [20] provide an analysis of M-estimators for RTSE with innovative outliers (IO), but show that the standard M-estimator has a break down point of *zero* in the presence of AO. This shows that standard M-estimators cannot handle even a non-zero fraction of corruptions. Recently, [8] proposed a method based on Least Trimmed Squares, which is closely related to our method, and used Monte Carlo simulations to validate the effectiveness of their method. [15] present a method based on robust filters in the more powerful ARMA model. Most of the estimators mentioned above are either not efficient (i.e. exponential time complexity) or do not provide non-asymptotic error rates. In contrast, we provide a consistent and nearly linear time algorithm that allows a large fraction of points to be corrupted. Recently, [3] studied time series with missing values but their results do not extend to cases with latent corruptions. Moreover, they consider the online setting as compared to the stochastic setting considered by our method.

Robust Regression: The goal in RLSR is to recover a parameter using noisy linear observations that are corrupted sparsely. RLSR is a classical problem in statistics, but computationally efficient, provable algorithms have been proposed only in recent years. The Least Trimmed Squares (LTS) method guarantees consistency but in general requires exponential running time [17, 1, 2]. Recently [22, 16] proposed L_1 norm minimization based methods for RLSR but their analyses do not guarantee consistent estimates in presence of dense unbiased i.i.d. noise. Recently, [5] proposed a hard thresholding style algorithm for RLSR but are unable to guarantee better than $O(\sigma)$ error in the estimation of $\mathbf{w}^*$ where σ is the standard deviation of noise. However, as detailed in section 3, their results holds in a weaker adversarial model than ours. In contrast, we provide nearly optimal $\sigma\frac{\sqrt{d}}{\sqrt{n}}$ error rates for our algorithm. [7] considers a stronger model where along with the response variables, the covariates can also be corrupted. However, their result also do not provide consistency guarantees and they can only tolerate $k \leq n/\sqrt{d}$ corruptions.

3 Robust Least Squares Regression

We use robust least squares regression (RLSR) as a warm up problem to introduce the tools, as well as establish notation that will be used for our time-series analysis. We present the problem formulation, propose our CRR algorithm, and then prove its consistency and robustness guarantees.

Problem Formulation and Notation: We are given a set of n data points $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ are the *covariates*, $\mathbf{y} \in \mathbb{R}^n$ is the vector of *responses* generated as

$$\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}, \tag{1}$$

for some *true* underlying model $\mathbf{w}^* \in \mathbb{R}^d$. The responses suffer two kinds of perturbations – *dense white noise* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ that is chosen in an i.i.d. fashion independently of the data X and the model $\mathbf{w}^*$, and *sparse adversarial corruptions* in the form of $\mathbf{b}$ whose support is chosen independently of $X, \mathbf{w}^*$ and $\boldsymbol{\epsilon}$. We assume that $\mathbf{b}^*$ is a k^*-sparse vector albeit one with potentially unbounded entries. The constant k^* will be called the *corruption index* of the problem. The above model is stronger than that of [5] which considers a fully adaptive adversary. However, whereas [5] is unable to give a consistent estimate, we give an algorithm CRR that does provide a consistent estimate. We also note that [5] is unable to give consistent estimates even in our model. As noted in the next section, our result requires significantly more fine analysis; standard ℓ_2-norm style anlaysis by [5] seems unlikely to lead to a consistency result in the robust regression setting.

We will require the notions of *Subset Strong Convexity* and *Subset Strong Smoothness* similar to [5] and reproduce the same below. For any set $S \subset [n]$, let $X_S := [\mathbf{x}_i]_{i\in S} \in \mathbb{R}^{p\times|S|}$ denote the matrix with columns in that set. We define $\mathbf{v}_S$ for a vector $\mathbf{v} \in \mathbb{R}^n$ similarly. $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ will denote, respectively, the smallest and largest eigenvalues of a square symmetric matrix X.

Definition 1 (SSC and SSS Properties)**.** *A matrix $X \in \mathbb{R}^{p\times n}$ is said to satisfy the* Subset Strong Convexity Property *(resp.* Subset Strong Smoothness Property*) at level k with strong convexity constant λ_k (resp. strong smoothness constant Λ_k) if the following holds:*

$$\lambda_k \leq \min_{|S|=k} \lambda_{\min}(X_S X_S^\top) \leq \max_{|S|=k} \lambda_{\max}(X_S X_S^\top) \leq \Lambda_k.$$

We refer the reader to the appendix for SSC/SSS bounds for Gaussian ensembles.

3.1 CRR: A Hard Thresholding Approach to Consistent Robust Regression

We now present our consistent method CRR for the RLSR problem. CRR takes a significantly different approach to the problem than previous works. Instead of attempting to exclude data points deemed unclean, CRR concentrates on correcting the errors instead. This allows CRR to work with the entire data set at all times, as opposed Torrent [5] that work with a fraction of the data.

Starting with the RLSR formulation $\min_{\mathbf{w}\in\mathbb{R}^p, \|\mathbf{b}\|_0\leq k^*} \frac{1}{2}\left\|X^\top\mathbf{w} - (\mathbf{y}-\mathbf{b})\right\|_2^2$, we realize that given any estimate $\widehat{\mathbf{b}}$ of the corruption vector, the optimal model with respect to this estimate is given by the expression

Algorithm 1 CRR: Consistent Robust Regression

Input: Covariates $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, \ldots, y_n]^\top$, corruption index k, tolerance ϵ

$\mathbf{b}^0 \leftarrow \mathbf{0}, t \leftarrow 0$,
$P_X \leftarrow X^\top (XX^\top)^{-1} X$
while $\left\|\mathbf{b}^t - \mathbf{b}^{t-1}\right\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(P_X \mathbf{b}^t + (I - P_X)\mathbf{y})$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}^t \leftarrow (XX^\top)^{-1} X(\mathbf{y} - \mathbf{b}^t)$

Algorithm 2 CRTSE: Consistent Robust Time Series Estimation

Input: Time-series data $y_i, i = -d+1, \ldots, n$, corruption index k, tolerance ϵ, time series order d, error trimming level $\widehat{\sigma}$

$y_i = \max\{\min\{y_i, \widehat{\sigma}\}, -\widehat{\sigma}\}$
$\mathbf{x}_i \leftarrow (y_{i-1}, \ldots, y_{i-d})^\top$, $X \leftarrow [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{y} \leftarrow (y_1, \ldots, y_n)^\top$, $P_X \leftarrow X^\top (XX^\top)^{-1} X$, $t \leftarrow 0$, $\mathbf{b}^0 \leftarrow 0$
while $\left\|\mathbf{b}^t - \mathbf{b}^{t-1}\right\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k^{\mathcal{G}}\left(P_X \mathbf{b}^t + (I - P_X)\,\mathbf{y}\right)$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}^t \leftarrow (XX^\top)^{-1} X(\mathbf{y} - \mathbf{b}^t)$

$\widehat{\mathbf{w}} = (XX^\top)^{-1} X(\mathbf{y} - \widehat{\mathbf{b}})$. Plugging this expression for $\widehat{\mathbf{w}}$ into the formulation allows us to reformulate the RLSR problem.

$$\min_{\|\mathbf{b}\|_0 \le k^*} f(\mathbf{b}) = \frac{1}{2} \|(I - P_X)(\mathbf{y} - \mathbf{b})\|_2^2 \tag{2}$$

where $P_X = X^\top (XX^\top)^{-1} X$. This greatly simplifies the problem by casting it as a *sparse parameter estimation* problem instead of a data subset selection problem. CRR directly optimizes (2) by using a form of iterative hard thresholding. At each step, CRR performs the following update: $\mathbf{b}^{t+1} = \mathrm{HT}_k(\mathbf{b}^t - \nabla f(\mathbf{b}^t))$, where k is a parameter for CRR. Any value $k \ge k^*$ suffices to ensure convergence and consistency. The hard thresholding operator is defined below.

Definition 2 (Hard Thresholding)**.** *For any $\mathbf{v} \in \mathbb{R}^n$, let the permutation $\sigma_{\mathbf{v}} \in S_n$ order elements of $\mathbf{v}$ in descending order of their magnitudes. Then for any $k \le n$, we define the hard thresholding operator as $\widehat{\mathbf{v}} = HT_k(\mathbf{v})$ where $\widehat{\mathbf{v}}_i = \mathbf{v}_i$ if $\sigma_{\mathbf{v}}^{-1}(i) \le k$ and 0 otherwise.*

We note that CRR functions with a fixed, unit step length, which is convenient in practice as it avoids step length tuning, something most IHT algorithms [9, 10] require. For the RLSR problem, we will consider data sets that are Gaussian ensembles i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Since CRR interacts with the data only using the projection matrix P_X, one can assume , without loss of generality, that the data points are generated from a standard Gaussian i.e. $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$. Our analysis will take care of the condition number of the data ensemble whenever it is apparent.

3.2 Convergence and Consistency Guarantees

Theorem 3. *Let $x_i \in \mathbb{R}^d, 1 \le i \le n$ be generated i.i.d. from a Gaussian distribution and let y_i's be generated using* (1) *for a fixed $\mathbf{w}^*$ and let σ^2 be the noise variance. Let the number of corruptions k^* be s.t. $k^* \le k \le n/10000$. Then, with probability at least $1 - \delta$,* CRR*, after $\mathcal{O}(\log(\|\mathbf{b}^*\|_2 / n) + \log(n/(\sigma \cdot d)))$ steps, ensures that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \mathcal{O}(\sigma\sqrt{d/n \log(nd/\delta)})$.*

The above result establishes consistency of the CRR method with $\tilde{\mathcal{O}}(\sigma\sqrt{d/n})$ error rates that are known to be statistically optimal, notably in the presence of gross and unbounded outliers. We reiterate that to the best of our knowledge, this is the first instance of a poly-time algorithm being shown to be consistent for the RLSR problem. It is also notable that the result allows the corruption index to be $k^* = \Omega(n)$, i.e. allows upto a *constant* factor of the total number of data points to be arbitrarily corrupted, while ensuring consistency, which existing results [5, 16] do not ensure.

For our analysis, we will divide CRR's execution into two phases – a *coarse convergence* phase and a *fine convergence* phase. CRR will enjoy a linear rate of convergence in both phases. However, the coarse convergence analysis will only ensure $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \mathcal{O}(\sigma)$. The fine convergence phase will then use a much more careful analysis of the algorithm to show that in at most $\mathcal{O}(\log n)$ more iterations, CRR

ensures $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \tilde{\mathcal{O}}(\sigma\sqrt{d/n})$, thus establishing consistency of the method. Existing methods, including TORRENT, are able to reach an error level $\mathcal{O}(\sigma)$, but no further.

Let $\boldsymbol{\lambda}^t := (XX^\top)^{-1}X(\mathbf{b}^t - \mathbf{b}^*)$, $\mathbf{g} := (I - P_X)\boldsymbol{\epsilon}$, and $\mathbf{v}^t = X^\top\boldsymbol{\lambda}^t + \mathbf{g}$. Let $S^* := \text{supp}(\mathbf{b}^*)$ true locations of the corruptions and $I^t := \text{supp}(\mathbf{b}^t) \cup \text{supp}(\mathbf{b}^*)$. Let $\text{MD}^t = \text{supp}(\mathbf{b}^*)\backslash\text{supp}(\mathbf{b}^t)$, $\text{FA}^t = \text{supp}(\mathbf{b}^t)\backslash\text{supp}(\mathbf{b}^*)$, and $\text{CI}^t = \text{supp}(\mathbf{b}^t) \cap \text{supp}(\mathbf{b}^*)$ respectively denote the coordinates that were *missed detections*, *false alarms*, and *correctly identifications*.

Coarse convergence: Here we establish a result that guarantees that after a certain number of steps T_0, CRR identifies the corruption vector with a relatively high accuracy i.e. $\left\|\mathbf{w}^{T_0} - \mathbf{w}^*\right\|_2 \leq \mathcal{O}(\sigma)$.

Lemma 4. *For any data matrix X that satisfies the SSC and SSS properties such that $\frac{2\Lambda_{k+k^*}}{\lambda_n} < 1$, CRR, when executed with a parameter $k \geq k^*$, ensures that after $T_0 = \mathcal{O}\left(\log \frac{\|\mathbf{b}^*\|_2}{\sqrt{n}}\right)$ steps, $\left\|\mathbf{b}^{T_0} - \mathbf{b}^*\right\|_2 \leq 3e_0$, where $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}}\right)$ for standard Gaussian designs.*

Using Lemma 17 (see the appendix), we can translate the above result to show that $\left\|\mathbf{w}^{T_0} - \mathbf{w}^*\right\|_2 \leq 0.95\sigma$, assuming $k = k^* \leq \frac{n}{150}$. However, Lemma 4 will be more useful in the following analysis.

Fine convergence: We now show that CRR progresses further at a linear rate to achieve a consistent solution. First Lemma 5 will show that $\left\|\boldsymbol{\lambda}^t\right\|_2$ can be bounded, apart from diminishing or negligible terms, by the amount of mass that is present in the false alarm coordinates MD^t. Lemma 6 will next bound this quantity. For all analyses hereon, we will assume $t > T_0$.

Lemma 5. *Suppose $k^* \leq k \leq n/10000$. Then with probability $1 - \delta$, at every time instant $t > T_0$, CRR ensures that $\|\boldsymbol{\lambda}^{t+1}\|_2 \leq \frac{1}{100}\|\boldsymbol{\lambda}^t\|_2 + 2\sigma\sqrt{\frac{2d}{n}\log\frac{d}{\delta}} + \frac{2.001}{\lambda_n}\|X_{FA^{t+1}}(X^\top_{FA^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{FA^{t+1}})\|_2$.*

We note that in the RHS above, the first term diminishes at a linear rate and the second term is a negligible quantity since it is $\tilde{\mathcal{O}}(\sqrt{d/n})$. In the following we bound the third term.

Lemma 6. *For $k^* \leq k \leq n/10000$, with probability at least $1-\delta$, CRR ensures at all $t > T_0$, $\frac{2.001}{\lambda_n}\left\|X_{FA^{t+1}}(X^\top_{FA^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{FA^{t+1}}\right.$ $0.98\left\|\boldsymbol{\lambda}^t\right\|_2 + C \cdot \sigma\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}$ for some constant C.*

Putting all these results together establishes Theorem 3. See Appendix A for a detailed proof.

4 Robust Time Series Estimation

Similar to RLSR, we formulate the Robust Time Series Estimation (RTSE) with additive outliers (AO) problem, propose our CRTSE algorithm, and prove its consistency and robustness guarantees.

Problem Formulation and Notation: Let $(x_{-d+1}, \ldots, x_n)$ be the "clean" time series which is a stationary and stable AR(d) process defined as $x_t = x_{t-1}\mathbf{w}_1^* + \cdots + x_{t-d}\mathbf{w}_d^* + \boldsymbol{\epsilon}_t$ where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. noise values chosen independently of the data and the model. We compactly represent this AR(d) process as,

$$\mathbf{y}^* = \overline{X}^\top\mathbf{w}^* + \boldsymbol{\epsilon},$$

where $\mathbf{y}^* = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, $\mathbf{x}_i = (x_{i-1}, \ldots, x_{i-d})^\top$, and $\overline{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. However, we do not observe the "clean" time series. Instead, we observe the time series $(y_{-d+1}, \ldots, y_n)$ which contains additive corruptions. Defining $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{d\times n}$ using $(y_{-d+1}, \ldots, y_n)$ in similar manner as $\mathbf{y}^*$ and $\overline{X}$ are defined using $(x_{-d+1}, \ldots, x_n)$, we have the resulting AO model as follows:

$$\mathbf{y} = \mathbf{y}^* + \mathbf{e}^* = X^\top\mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{b}^*, \tag{3}$$

where $\mathbf{e}^*$ is the actual corruption vector (k^*-sparse), and $\mathbf{b}^*$ is the resulting model corruption vector (with at most k^*-blocks of size d being non-zero). See (19) (see Appendix B.2) for a clearer characterization of the $\mathbf{y}, X$.

Now, given $\mathbf{y}, X$, our goal will be to recover a consistent estimate of the parameter $\mathbf{w}^*$. For our results the following simple observation would be crucial: since $\mathrm{supp}(\mathbf{b}^*)$ is a union of k^* groups (intervals) of size d, we have $\|\mathbf{b}^*\|_0^{\mathcal{G}} \leq 2k^*$, where $\|\mathbf{b}\|_0^{\mathcal{G}}$ is the Group-ℓ_0 pseudo-norm of $\mathbf{b}$ that we define below. For a set of groups S, $\mathrm{supp}(S;\mathcal{G}) = \{G_i, i \in S\}$.

We now define certain quantities that are crucial in understanding the $\mathrm{AR}(d)$ process. The *spectral density* of the "clean" $\mathrm{AR}(d)$ process $\mathbf{y}^*$ is given by:

$$\rho_{\mathbf{w}^*}(\omega) = \frac{\sigma^2}{\left(1 - \sum_{k=1}^{d} \mathbf{w}_k^* e^{ik\omega}\right)\left(1 - \sum_{k=1}^{d} \mathbf{w}_k^* e^{-ik\omega}\right)}, \text{ for } \omega \in [0, 2\pi]. \tag{4}$$

We define $\mathcal{M}_{\mathbf{w}^*} := \sup_{\omega\in[0,2\pi]} \rho_{\mathbf{w}^*}(\omega)$ and $\mathfrak{m}_{\mathbf{w}^*} := \inf_{\omega\in[0,2\pi]} \rho_{\mathbf{w}^*}(\omega)$. Another constant $\mathcal{M}_W$ will also appear in our results (see Appendix B.2 for a brief primer on $\mathrm{AR}(d)$ process).

For our analysis, we will also require notions of *Sub-group Strong Convexity* and *Sub-group Strong Smoothness* for the time series which we define below. For any $k \leq \frac{n}{d}$, we let $\mathcal{S}_k^{\mathcal{G}} = \left\{\mathrm{supp}(S;\mathcal{G}) : S \subseteq \left[\frac{n}{d}\right] \text{ s.t. } |S| = k\right\}$ denote the *set of all* collections of k groups from $\mathcal{G}$.

Definition 7 (SGSC/SGSS)**.** *A matrix $X \in \mathbb{R}^{d\times n}$ satisfies the Subgroup Strong Convexity Property (resp. Subgroup Strong Smoothness Property) at level k with strong convexity constant λ_k (resp. strong smoothness constant Λ_k) if the following holds:*

$$\lambda_k \;\leq\; \min_{S\in\mathcal{S}_k^{\mathcal{G}}} \lambda_{\min}\left(X_S X_S^\top\right) \;\leq\; \max_{S\in\mathcal{S}_k^{\mathcal{G}}} \lambda_{\max}\left(X_S X_S^\top\right) \;\leq\; \Lambda_k.$$

4.1 CRTSE: A Block Sparse Hard Thresholding Approach to Consistent Robust Time Series Estimation

We now present our CRTSE method for obtaining consistent estimates in the RTSE problem. By following the similar approach as CRR, we begin with the RTSE formulation $\min_{\mathbf{w}\in\mathbb{R}^d, \|\mathbf{b}\|_0^{\mathcal{G}}\leq k^*} \frac{1}{2}\left\|X^\top\mathbf{w} - (\mathbf{y}-\mathbf{b})\right\|_2^2$, and observe that for any given estimate $\widehat{\mathbf{b}}$ of the corruption vector, the optimal model with respect to that estimate is $\widehat{\mathbf{w}} = (XX^\top)^{-1}X(\mathbf{y}-\widehat{\mathbf{b}})$. Then by plugging this expression for $\widehat{\mathbf{w}}$ into the formulation, we reformulate the RTSE problem as follows

$$\min_{\|\mathbf{b}\|_0^{\mathcal{G}}\leq k^*} f(\mathbf{b}) = \frac{1}{2}\|(I - P_X)(\mathbf{y}-\mathbf{b})\|_2^2 \tag{5}$$

where $P_X = X^\top(XX^\top)^{-1}X$. CRTSE uses a variant of iterative hard thresholding to optimize the above formulation. At every iteration, CRTSE takes a step along the negative gradient of the function f and then performs group hard thresholding to select the *top* k aligned groups (i.e. groups in $\mathcal{G}$) of the resulting vector and setting the rest to zero.

$$\mathbf{b}^{t+1} = \mathrm{HT}_k^{\mathcal{G}}(\mathbf{b}^t - \nabla f(\mathbf{b}^t)),$$

where $k \geq 2k^*$ and the group hard thresholding operator is defined below.

Definition 8 (Group Hard Thresholding)**.** *For any vector $\mathbf{g} \in \mathbb{R}^n$, let $\sigma_{\mathbf{g}} \in S_{\frac{n}{d}}$ be the permutation s.t. $\sum_{j\in G_{\sigma_{\mathbf{g}}(1)}} |\mathbf{g}_j|^2 \geq \sum_{j\in G_{\sigma_{\mathbf{g}}(2)}} |\mathbf{g}_j|^2 \geq \ldots \geq \sum_{j\in G_{\sigma_{\mathbf{g}}(\frac{n}{d})}} |\mathbf{g}_j|^2$. Then for any $k \leq \frac{n}{d}$, we define the group hard thresholding operator as $\widehat{\mathbf{g}} = HT_k^{\mathcal{G}}(\mathbf{g})$ where*

$$\widehat{\mathbf{g}}_i \;=\; \begin{cases} \mathbf{g}_i & \text{if } \sigma_{\mathbf{g}}^{-1}(\lceil \frac{i}{d} \rceil) \leq k \\ 0 & \text{else} \end{cases}$$

We note that this step can be done in quasi linear time. Due to the delicate correlations between data points in the time series, in order to keep the problem well conditioned (see Theorem 22 and Remark 23),

we will perform a pre-processing step on the corrupted time series instances $y_i, i = -d+1, \ldots, n$ as follows: $y_i = \max\{\min\{y_i, \widehat{\sigma}\}, -\widehat{\sigma}\}$, where $\widehat{\sigma} = \mathcal{O}\left(\sqrt{\log n}\sigma\right)$. Note that since the clean underlying time series is a Gaussian process $\boldsymbol{\epsilon}_i \leq \mathcal{O}\left(\sigma\sqrt{\log n}\right)$ and all its entries are, with high probability, bounded by $\widehat{\sigma}$. Thus we will not clip any clean point because of the above step but ensure that we can, from now on, assume that $\|\mathbf{b}^*\|_\infty \leq \widehat{\sigma}$.

4.2 Convergence and Consistency Guarantees

We now present the estimation error bound for our CRTSE algorithm.

Theorem 9. *Let* $\mathbf{y}$ *be generated using AR*(d) *process with* k^* *additive outliers (see* (3)*). Also, let* $k^* \leq k \leq C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$ *(for some universal constant* $C > 0$*). Then, with probability at least* $1-\delta$, CRTSE*, after* $\mathcal{O}(\log(\|\mathbf{b}^*\|_2/n) + \log(n/(\sigma\cdot d)))$ *steps, ensures that* $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \mathcal{O}\left(\sigma\mathcal{M}_{\mathbf{w}^*}/\mathfrak{m}_{\mathbf{w}^*}\sqrt{d\log n/n\log(d/\delta)}\right)$.

The result does establish consistency of the CRTSE method as it offers convergence to $\widetilde{\mathcal{O}}\left(\sigma\sqrt{d\log n/n}\right)$ error levels. Also note that in typical time series data, d lies in the range $5-10$. As in the case of CRR, this is the first instance of a poly-time algorithm being shown to be consistent for the RTSE problem.

Following the similar approach of the consistency analysis for CRR, we will first ensure that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \mathcal{O}(\sigma)$. Then in the fine analysis phase, we will show that after additional $\mathcal{O}(\log n)$ iterations, CRTSE ensures $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \widetilde{\mathcal{O}}\left(\sigma\sqrt{d\log n/n}\right)$.

Coarse convergence: Here we establish a result that after a certain number of iterations, CRTSE identifies the corruption vector with a relatively high accuracy. Our analysis relies on a novel Theorem 22, which is a *key result* that shows that the AR(d) process with AO indeed satisfies SGSC and SGSS properties (see Definition 8), as long as the number of corruptions k^* is small.

Theorem 10. *For any data matrix* X *that satisfies the SGSC and SGSS properties such that* $4\Lambda_{k+k^*} < \lambda_{\frac{n}{d}}$, CRTSE*, when executed with a parameter* $k \geq k^*$*, ensures that after* $T_0 = \mathcal{O}(\log(\|\mathbf{b}^*\|_2/\sqrt{n}))$ *steps,* $\left\|\mathbf{b}^{T_0} - \mathbf{b}^*\right\|_2 \leq 5e_0$*. Additionally, if* X *is generated using our AR*(d) *process with AO (see* (3)*), then* $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)d\log\frac{n}{\delta(k+k^*)d}}\right)$.

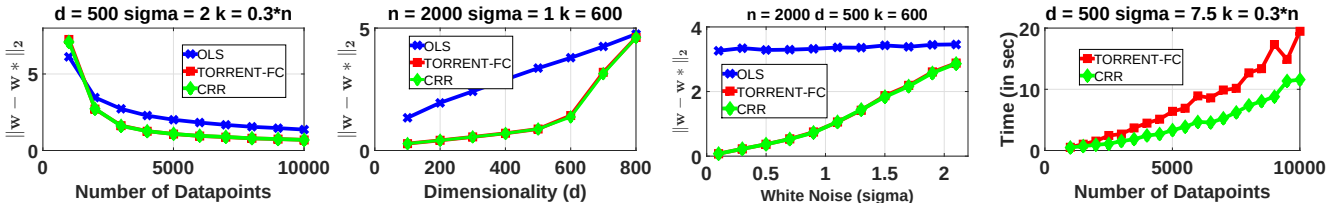
Note that if X is given by AR(d) process with AO model and if k is sufficiently small i.e. $k^* \leq k \leq C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$ (for some universal constant $C > 0$) and n is sufficiently large enough, then with probability at least $1-\delta$, we have $4\Lambda_{k+k^*} < \lambda_{\frac{n}{d}}$. See Remark 23 for more details.

Fine Convergence: As was the case in least squares regression, we will now sketch a proof that the CRTSE algorithm indeed moves beyond the convergence level achieved in the coarse analysis and proceeds towards a consistent solution at a linear rate. We begin by noting that by applying Lemma 24, we can derive a result similar to Lemma 17. With high probability, we have for all $t > 1$

$$\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq C\cdot\frac{\Lambda_n}{\lambda_n}\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{d}{\delta}} + \left\|\boldsymbol{\lambda}^t\right\|_2\right), \tag{6}$$

for a universal constant C. We note that for large enough n, Lemma 19 shows that $\frac{\Lambda_n}{\lambda_n} = \mathcal{O}(1)$. Since the first term in the bracket is a negligible term, one that does not hinder consistency, save log factors, we are just left to establish the convergence of the iterates $\boldsymbol{\lambda}^t$. We next note that Lemma 24, along with the fact that the locations of the corruptions were decided obliviously and independently of the noise values $\{\epsilon_i\}$, allows us to also prove the following equivalent of Lemma 5 for the time series case as well: with high probability, for every time instant $t > T_0$, we have

$$\left\|\boldsymbol{\lambda}^{t+1}\right\|_2 \leq \frac{1}{100}\left\|\boldsymbol{\lambda}^t\right\|_2 + C\cdot\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{d}{\delta}} + \frac{1}{\lambda_n}\left(1+\frac{\Lambda_n}{\lambda_n}\right)\left\|X_{\mathrm{FA}^{t+1}}(X_{\mathrm{FA}^{t+1}}^\top\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2\right), \tag{7}$$



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

Figure 1: (a), (b) and (c) show variation of recovery error with varying n, d and σ. CRR and TORRENT show better recovery properties than the non-robust OLS. These plots also ascertain the $\sqrt{n}$-consistency of CRR as is shown in the theoretical analysis. (d) shows the average CPU run time of TORRENT and CRR with increasing sample size. CRR can be upto 2x faster than TORRENT while ensuring similar recovery properties.

for some universal constant C. Noticing yet again that $\frac{\Lambda_n}{\lambda_n} = \mathcal{O}(1)$ leaves us to prove a bound on the quantity $\left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2$. We now notice that one can upper bound this quantity by $\left\|X_{\mathrm{FA}^{t+1}}(X^\top_{S^t_k}\boldsymbol{\lambda}^t + \mathbf{g}_{S^t_k})\right\|_2$ by selecting the set S^t_k of the top k elements by magnitude in the vector $X^\top_{\overline{S^*}}\boldsymbol{\lambda}^t + \mathbf{g}_{\overline{S^*}}$. This allows us to establish the following result.

Lemma 11. *Suppose $k^* \le k \le n/(C'\rho(\mathbf{w}^*)d\log n)$ for some large enough constant C'. Then with probability at least $1-\delta$,* CRR *ensures at every time instant $t > T_0$*

$$\frac{C}{\lambda_n}\left(1+\frac{\Lambda_n}{\lambda_n}\right)\left\|X_{FA^{t+1}}(X^\top_{FA^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{FA^{t+1}})\right\|_2 \le 0.5\left\|\boldsymbol{\lambda}^t\right\|_2 + \mathcal{O}\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{1}{\delta}}\right)$$

Above lemma with (7) suffices to establish Theorem 9. See Appendix B for details of all the steps sketched above.

5 Experiments

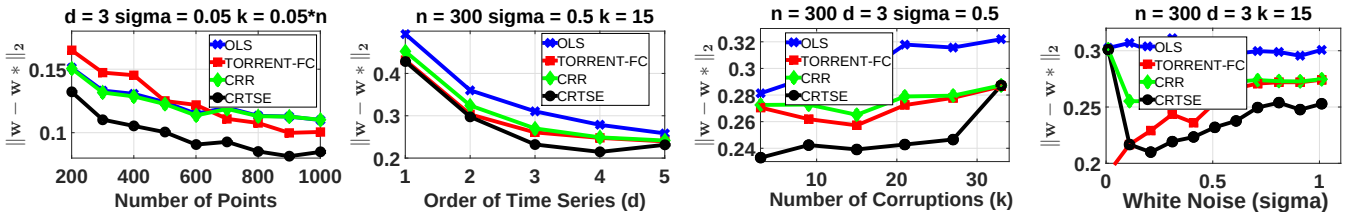
Several numerical simulations were carried out on synthetically generated linear regression and AR(d) time-series data with outliers. The experiments show that in the robust linear regression setting, CRR gives a consistent estimator and is 2x times faster as compared with TORRENT [5] while in the robust AR(d) time-series setting, CRTSE gives a consistent estimator and offers statistically better recovery properties as compared with baseline algorithms.

5.1 Robust Linear Regression

Data: For the RLSR problem, the regressor $\mathbf{w}^* \in \mathbb{R}^d$ was chosen to be a random unit norm vector. The data matrix was generated as each $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$. The k^* non-zero locations of the corruption vector $\mathbf{b}^*$ were chosen uniformly at random from $[n]$ and the value of the corruptions were set to $b^*_i \sim U(10, 20)$. The response variables $\mathbf{y}$ were then generated as $y_i = \langle \mathbf{x}_i, \mathbf{w}^*\rangle + \eta_i + b^*_i$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$. All plots for the RLSR problem have been generated by averaging the results over 20 random instances of the data and regressor.

Baseline Algorithms: We compare CRR with two baseline algorithms: Ordinary Least Squares (OLS) and TORRENT ([5]). All the three algorithms were implemented in Matlab and were run on a single core 2.4GHz machine with 8GB RAM.

Recovery Properties & Timing: As can be observed from Figure(1), CRR performs as well as TORRENT in terms of the residual error $\|\mathbf{w}-\mathbf{w}^*\|_2$ and both their performances are better as compared with the non-robust OLS method. Further, figures 1(a), 1(b) and 1(c) explain our near optimal recovery bound of $\sigma\sqrt{\frac{d}{n}}$ by showing the corresponding variation of the recovery error with variations in n, d and σ, respectively. Figure 1(d) shows a comparison of variation of average CPU time (in secs) with increasing number of data samples and shows that CRR can be upto 2x faster than TORRENT while provably guaranteeing consistent estimates for the regressor.



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

Figure 2: (a), (b), (c), and (d) show variation of recovery error with varying n, d, k and σ, respectively. CRTSE outperforms OLS, and both the point-wise thresholding algorithms, TORRENT and CRR. Also, the decreasing error with increasing n shows the consistency of our estimator in this regime.

5.2 Robust Time Series with Additive Corruptions

Data: For the RTSE problem, the regressor $\mathbf{w}^* \in \mathbb{R}^d$ was chosen to be a random vector with $O(\frac{1}{\sqrt{d}})$ norm (to avoid the time-series from diverging). The initial d points of the time-series are chosen as $x_i \sim \mathcal{N}(0,1)$ for $i = 1 \dots d$. The time-series, generated according AR (d) model with regressor $\mathbf{w}^*$, was then allowed to stabilize for the next 100 time-steps. We consider the points generated in the next n time steps as x_i for $i = 1 \dots n$. The k^* non-zero locations of the corruption vector $\mathbf{b}^*$ were chosen uniformly at random from $[n]$ and the value of the corruptions were set to $b_i^* \sim U(10, 20)$. The observed time series is then generated as $y_i = x_i + b_i^*$. All plots for the RTSE problem have been generated by averaging the outcomes over 200 random runs of the above procedure.

Baseline Algorithms: We compare CRTSE with three baseline algorithms: Ordinary Least Squares (OLS) , TORRENT ([5]) and CRR. For TORRENT and CRR, we set the thresholding parameter $k = 2k^*d$ and compare results with CRTSE. All simulations were done on a single core 2.4GHz machine with 8GB RAM.

Recovery Properties: Figure 2 shows the variation of recovery error $\|\mathbf{w} - \mathbf{w}^*\|_2$ for the AR (d) time-series with Additive Corruptions. CRTSE outperforms all three competitor baselines: OLS, TORRENT and CRR. Since CRTSE uses a group thresholding based algorithm as compared with TORRENT and CRR which use point-wise thresholding, CRTSE is able to identify blocks which contain both response and data corruptions and give better estimates for the regressor. Also, figure 2(a) shows that the recovery error goes down with increasing number of points in the time-series, as is evident from our consistency analysis of CRTSE.

References

[1] Jan Ámos Višek. The least trimmed squares. Part I: Consistency. *Kybernetika*, 42:1–36, 2006.

[2] Jan Ámos Višek. The least trimmed squares. Part II: $\sqrt{n}$-consistency. *Kybernetika*, 42:181–202, 2006.

[3] Oren Anava, Elad Hazan, and Assaf Zeevi. Online time series prediction with missing data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2191–2199, 2015.

[4] Sumanta Basu and George Michailidis. Regularized Estimation in Sparse High-dimensional Time Series Models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

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

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

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

[8] Christophe Croux and Kristel Joossens. Robust estimation of the vector autoregressive model by a least trimmed squares procedure. In *COMPSTAT 2008*, pages 489–501. Springer, 2008.

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

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

[11] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust Statistics: Theory and Methods*. J. Wiley, 2006.

[12] R. Douglas Martin. Robust estimation for time series autoregressions. In ROBERT L. LAUNER and GRAHAM N. WILKINSON, editors, *Robustness in Statistics*, pages 147 – 176. Academic Press, 1979.

[13] R. Douglas Martin and Judy Zeh. Robust generalized m-estimates for autoregressive parameters: smallsample behavior and applications. Technical Report 214, University of Washington, Seattle, 1978.

[14] Igor Melnyk and Arindam Banerjee. Estimating structured vector autoregressive model. arXiv:1602.06606 (math.ST), 2016.

[15] Nora Muler, Daniel Pena, and Victor J. Yohai. Robust estimation for arma models. *The Annals of Statistics*, 37(2):816–840, 2009.

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

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

[18] Mark Rudelson and Roman Vershynin. Hanson-Wright Inequality and Sub-gaussian Concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013.

[19] Ohad Shamir. A variant of azuma's inequality for martingales with subgaussian tails. arXiv:1110.2392 (cs.LG), 2011.

[20] Norbert Stockinger and Rudolf Dutter. Robust time series analysis: A survey. *Kybernetika*, 23(7):1–3, 1987.

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

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

[23] Tong Zhang. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Trans. Inf. Theory*, 57:4689–4708, 2011.

A Supplementary Material for Consistent Robust Regression

A.1 SSC/SSS guarantees

In this section we restate some results from [5] which are required for the convergence analysis of the RLSR problem.

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

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

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

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

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

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

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

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

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

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

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

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

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

A.2 Convergence Proofs for CRR

Theorem 3. *For $k^* \leq k \leq n/10000$ and Gaussian designs, with probability at least $1-\delta$, CRR, after $\mathcal{O}\left(\log \frac{\|\mathbf{b}^*\|_2}{n} + \log \frac{n}{d}\right)$ steps, ensures that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \mathcal{O}\left(\frac{\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}\right)$.*

Proof. Putting Lemmata 5 and 6 establishes that

$$\left\|\boldsymbol{\lambda}^{t+1}\right\|_2 \leq 0.99\left\|\boldsymbol{\lambda}^t\right\|_2 + C\sigma\sqrt{\frac{d}{n}\log\frac{nd}{\delta}},$$

which ensures a linear convergence of the terms $\left\|\boldsymbol{\lambda}^t\right\|_2$ to a value $\mathcal{O}\left(\sigma\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}\right)$. Applying Lemma 17 then finishes off the result. □

Lemma 4. *For any data matrix X that satisfies the SSC and SSS properties such that $\frac{2\Lambda_{k+k^*}}{\lambda_n} < 1$, CRR, when executed with a parameter $k \geq k^*$, ensures that after $T_0 = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{\sqrt{n}}\right)$ steps, $\left\|\mathbf{b}^{T_0} - \mathbf{b}^*\right\|_2 \leq 3e_0$, where $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}}\right)$ for standard Gaussian designs.*

Proof. We start with the update step in CRR, and use the fact that $\mathbf{y} = X^\top\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}$ to rewrite the update as

$$\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(P_X\mathbf{b}^t + (I-P_X)(X^\top\mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon})).$$

Since $X^\top = P_X X^\top$, we get, using the notation set up before,

$$\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(\mathbf{b}^* + X^\top\boldsymbol{\lambda}^t + \mathbf{g}).$$

Since $k \geq k^*$, using the properties of the hard thresholding step gives us

$$\left\|\mathbf{b}^{t+1}_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}})\right\|_2 \leq \left\|\mathbf{b}^*_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}})\right\|_2 = \left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}}\right\|_2.$$

This, upon applying the triangle inequality, gives us

$$\left\|\mathbf{b}^{t+1} - \mathbf{b}^*\right\|_2 \leq 2\left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}}\right\|_2.$$

Now, using the SSC and SSS properties of X, we can show that $\left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t\right\|_2 = \left\|X^\top_{I^{t+1}}(XX^\top)^{-1}X^\top_{I^t}(\mathbf{b}^t - \mathbf{b}^*)\right\|_2 \leq \frac{\Lambda_{k+k^*}}{\lambda_n}\left\|\mathbf{b}^t - \mathbf{b}^*\right\|_2$.

Since $\boldsymbol{\epsilon}$ is a Gaussian vector, using tail bounds for Chi-squared random variables (for example, see [5, Lemma 20]), for any set S of size $k+k^*$, we have with probability at least $1-\delta$,$\|\boldsymbol{\epsilon}_S\|_2^2 \leq \sigma^2(k+k^*) + 2e\sigma^2\sqrt{6(k+k^*)\log\frac{1}{\delta}}$. Taking a union bound over all sets of size $(k+k^*)$ and $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ gives us, with probability at least $1-\delta$, for all sets S of size at most $(k+k^*)$,

$$\|\boldsymbol{\epsilon}_S\|_2 \leq \sigma\sqrt{(k+k^*)}\sqrt{1 + 2e\sqrt{6\log\frac{en}{\delta(k+k^*)}}}$$

Using tail bounds on Gaussian random variables[1], we can also show that for every i, with probability at least $1-\delta$, we have $\|(X\epsilon)_i\|_2 \leq \sigma\left\|(X^\top)_i\right\|_2\sqrt{2\log\frac{1}{\delta}}$. Taking a union bound gives us, with the same confidence, $\|X\boldsymbol{\epsilon}\|_2^2 \leq 2\sigma^2\|X\|_F^2\log\frac{d}{\delta} \leq 2\sigma^2 d\Lambda_n\log\frac{d}{\delta}$. This allows us to bound $\|\mathbf{g}_{I^{t+1}}\|_2$

$$\|\mathbf{g}_{I^{t+1}}\|_2 = \left\|\boldsymbol{\epsilon}_{I^{t+1}} - X^\top_{I^{t+1}}(XX^\top)^{-1}X\boldsymbol{\epsilon}\right\|_2$$

[1] $\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt \leq \frac{1}{\sqrt{2\pi}}\int_x^\infty \frac{t}{x}e^{-t^2/2}dt = \frac{1}{x\sqrt{2\pi}}e^{-x^2/2}$

$$
\begin{aligned}
&\leq \sigma\sqrt{(k+k^*)}\sqrt{1+2e\sqrt{6\log\frac{en}{\delta(k+k^*)}}} + \sigma\frac{\sqrt{\Lambda_{k+k^*}\Lambda_n}}{\lambda_n}\sqrt{2d\log\frac{d}{\delta}} \\
&\leq \underbrace{\sigma\sqrt{(k+k^*)}\sqrt{1+2e\sqrt{6\log\frac{en}{\delta(k+k^*)}}}}_{e_0}\left(1+\sqrt{\frac{2d}{n}\log\frac{d}{\delta}}\right) \\
&= 1.0003e_0,
\end{aligned}
$$

where the second last step is true for Gaussian designs and sufficiently large enough n. Note that e_0 does note depend on the iterates and is thus, a constant. This gives us

$$
\left\|\mathbf{b}^{t+1}-\mathbf{b}^*\right\|_2 \leq \frac{2\Lambda_{k+k^*}}{\lambda_n}\left\|\mathbf{b}^t-\mathbf{b}^*\right\|_2 + 2.0006e_0.
$$

For data matrices sampled from Gaussian ensembles, whose SSC and SSS properties will be established later, assuming $n \geq d\log d$, we have $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}}\right)$. Thus, if $\frac{2\Lambda_{k+k^*}}{\lambda_n} < 1$, then in $T_0 = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{e_0}\right) = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{\sqrt{n}}\right)$ steps, CRR ensures that $\left\|\mathbf{b}^{T_0}-\mathbf{b}^*\right\|_2 \leq 2.0009e_0$. □

Lemma 17. *Let $\lambda_{\min}(\Sigma)$ be the smallest eigenvalue of the covariance matrix of the distribution $\mathcal{N}(\mathbf{0},\Sigma)$ that generates the data points. Then at any time instant t, we have $\|\mathbf{w}^t-\mathbf{w}^*\|_2 \leq \frac{2}{\lambda_{\min}(\Sigma)}\left(2\sigma\sqrt{\frac{d}{n}\log\frac{d}{\delta}} + \left\|\boldsymbol{\lambda}^t\right\|_2\right)$.*

Proof. As described in Algorithm 1, $\mathbf{w}^t = (XX^\top)^{-1}X(\mathbf{y}-\mathbf{b}^t) = \mathbf{w}^* + (XX^\top)^{-1}X(\boldsymbol{\epsilon}+\mathbf{b}^*-\mathbf{b}^t)$. Thus, we get

$$
\begin{aligned}
\left\|\mathbf{w}^t-\mathbf{w}^*\right\|_2 &\leq \frac{1}{\lambda_{\min}(XX^\top)}\left\|X^\top(\mathbf{w}^t-\mathbf{w}^*)\right\|_2 \\
&\leq \frac{1}{n\lambda_{\min}(\Sigma)-C_\Sigma\sqrt{n}}\left\|X^\top(\mathbf{w}^t-\mathbf{w}^*)\right\|_2 \\
&\leq \frac{1}{n\lambda_{\min}(\Sigma)-C_\Sigma\sqrt{n}}\left\|\overline{X}^\top(\overline{XX}^\top)^{-1}\overline{X}(\boldsymbol{\epsilon}+\mathbf{b}^*-\mathbf{b}^t)\right\|_2 \\
&\leq \frac{\Lambda_n}{n\lambda_{\min}(\Sigma)-C_\Sigma\sqrt{n}}\left\|(\overline{XX}^\top)^{-1}\overline{X}(\boldsymbol{\epsilon}+\mathbf{b}^*-\mathbf{b}^t)\right\|_2 \\
&\leq \frac{2}{\lambda_{\min}(\Sigma)}\left(2\sigma\sqrt{\frac{d}{n}\log\frac{d}{\delta}} + \left\|\boldsymbol{\lambda}^t\right\|_2\right),
\end{aligned}
$$

where the second step follows from results on eigenvalue bounds for data matrices drawn from non-spherical Gaussians, where C_Σ is a constant dependent on the subGaussian norm of the distribution, and the last step assumes $n \geq \frac{2C_\Sigma}{\lambda_{\min}(\Sigma)}$ and uses the proof technique used in Lemma 4 to get

$$
\left\|(\overline{XX}^\top)^{-1}\overline{X}\boldsymbol{\epsilon}\right\|_2 \leq \sigma\frac{\sqrt{\Lambda_n}}{\lambda_n}\sqrt{2d\log\frac{d}{\delta}} \leq 2\sigma\sqrt{\frac{d}{n}\log\frac{d}{\delta}}.
$$

□

Lemma 5. *Suppose $k^* \leq k \leq n/10000$. Then with probability $1-\delta$, at every time instant $t > T_0$,* CRR *ensures that $\left\|\boldsymbol{\lambda}^{t+1}\right\|_2 \leq \frac{1}{100}\left\|\boldsymbol{\lambda}^t\right\|_2 + 2\sigma\sqrt{\frac{2d}{n}\log\frac{d}{\delta}} + \frac{2.001}{\lambda_n}\left\|X_{FA^{t+1}}(X_{FA^{t+1}}^\top\boldsymbol{\lambda}^t + \mathbf{g}_{FA^{t+1}})\right\|_2$.*

Proof. We have $\mathbf{b}^{t+1} = \mathrm{HT}_k(\mathbf{b}^* + X^\top\boldsymbol{\lambda}^t + \mathbf{g})$. To analyze $\boldsymbol{\lambda}^{t+1} = (XX^\top)^{-1}X(\mathbf{b}^{t+1}-\mathbf{b}^*)$, we start by looking at $X(\mathbf{b}^{t+1}-\mathbf{b}^*) = X_{\mathrm{MD}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{MD}^{t+1}}-\mathbf{b}^*_{\mathrm{MD}^{t+1}}) + X_{\mathrm{FA}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{FA}^{t+1}}-\mathbf{b}^*_{\mathrm{FA}^{t+1}}) + X_{\mathrm{CI}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{CI}^{t+1}}-\mathbf{b}^*_{\mathrm{CI}^{t+1}})$. We then have

$$
X_{\mathrm{MD}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{MD}^{t+1}}-\mathbf{b}^*_{\mathrm{MD}^{t+1}}) = X_{\mathrm{MD}^{t+1}}(-\mathbf{b}^*_{\mathrm{MD}^{t+1}})
$$

$$X_{\mathrm{CI}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{CI}^{t+1}} - \mathbf{b}^*_{\mathrm{CI}^{t+1}}) = X_{\mathrm{CI}^{t+1}}(X^\top_{\mathrm{CI}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{CI}^{t+1}})$$
$$X_{\mathrm{FA}^{t+1}}(\mathbf{b}^{t+1}_{\mathrm{FA}^{t+1}} - \mathbf{b}^*_{\mathrm{FA}^{t+1}}) = X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}}).$$

This gives us upon completing the terms, and using $\mathrm{CI}^{t+1} \uplus \mathrm{MD}^{t+1} = S^*$,

$$X(\mathbf{b}^{t+1} - \mathbf{b}^*) = X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}}) + X_{S^*}(X^\top_{S^*}\boldsymbol{\lambda}^t + \mathbf{g}_{S^*}) - X_{\mathrm{MD}^{t+1}}(\mathbf{b}^*_{\mathrm{MD}^{t+1}} + X^\top_{\mathrm{MD}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{MD}^{t+1}}).$$

Now due to the hard thresholding operation, we have $\left\|\mathbf{b}^*_{\mathrm{MD}^{t+1}} + X^\top_{\mathrm{MD}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{MD}^{t+1}}\right\|_2 \le \left\|X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}}\right\|_2$. This gives us

$$\begin{aligned}
\left\|X_{\mathrm{MD}^{t+1}}(\mathbf{b}^*_{\mathrm{MD}^{t+1}} + X^\top_{\mathrm{MD}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{MD}^{t+1}})\right\|_2 &= \left\|X(\mathbf{b}^*_{\mathrm{MD}^{t+1}} + X^\top_{\mathrm{MD}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{MD}^{t+1}})\right\|_2 \\
&\le \Lambda_n \left\|\mathbf{b}^*_{\mathrm{MD}^{t+1}} + X^\top_{\mathrm{MD}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{MD}^{t+1}}\right\|_2 \\
&\le \Lambda_n \left\|X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}}\right\|_2 \\
&\le \frac{\Lambda_n}{\lambda_n} \left\|X(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2 \\
&= \frac{\Lambda_n}{\lambda_n} \left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2 \\
&\le 1.001 \left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2,
\end{aligned}$$

where the last step uses a large enough n so that the data matrix X is well conditioned. Thus,

$$\begin{aligned}
\left\|\boldsymbol{\lambda}^{t+1}\right\|_2 &= \left\|(XX^\top)^{-1}X(\mathbf{b}^{t+1} - \mathbf{b}^*)\right\|_2 \\
&\le \frac{1}{\lambda_n}\left\|X_{S^*}(X^\top_{S^*}\boldsymbol{\lambda}^t + \mathbf{g}_{S^*})\right\|_2 + \frac{2.001}{\lambda_n}\left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2 \\
&\le \frac{1}{100}\left\|\boldsymbol{\lambda}^t\right\|_2 + 2\sigma\sqrt{\frac{2d}{n}\log\frac{d}{\delta}} + \frac{2.001}{\lambda_n}\left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2,
\end{aligned}$$

where the third step follows by observing that the columns of X are (statistically equivalent to) i.i.d. samples from a standard Gaussian, the fact that the support of the corruptions S^* is chosen independently of the data and the noise, and requiring that $k^* \le \frac{n}{100}$. □

Lemma 6. *Suppose $k^* \le k \le n/10000$. Then with probability at least $1-\delta$,* CRR *ensures at every time instant $t > T_0$, for some constant C*

$$\frac{2.001}{\lambda_n}\left\|X_{FA^{t+1}}(X^\top_{FA^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{FA^{t+1}})\right\|_2 \le 0.98\left\|\boldsymbol{\lambda}^t\right\|_2 + C\sigma\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}$$

Proof. For this we first observe that, since entries in the set FA^{t+1} survived the hard thresholding step, they must have been the largest elements by magnitude in the set $\overline{S^*}$ i.e.

$$X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}} = \mathrm{HT}_{\left|\mathrm{FA}^{t+1}\right|}(X^\top_{\overline{S^*}}\boldsymbol{\lambda}^t + \mathbf{g}_{\overline{S^*}})$$

Note that $\left|\mathrm{FA}^{t+1}\right| \le k$ and $\overline{S^*}$ is a fixed set of size $n-k^*$ with respect to the data points and the Gaussian noise. Thus, if we denote by S^t_k, the set of top k coordinates by magnitude in $\overline{S^*}$ i.e.

$$X^\top_{S^t_k} + \mathbf{g}_{S^t_k} = \mathrm{HT}_k(X^\top_{\overline{S^*}}\boldsymbol{\lambda}^t + \mathbf{g}_{\overline{S^*}}),$$

then $\left\|X_{\mathrm{FA}^{t+1}}(X^\top_{\mathrm{FA}^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{\mathrm{FA}^{t+1}})\right\|_2 \le \left\|X_{S^t_k}(X^\top_{S^t_k}\boldsymbol{\lambda}^t + \mathbf{g}_{S^t_k})\right\|_2$. Thus, all we need to do is bound this term. In the following, we will, for sake of simplicity, omit the subscript $\overline{S^*}$.

Before we move ahead, we make a small change to notation for convenience. At the moment, we are defining $\boldsymbol{\lambda}^t = (XX^\top)^{-1}(\mathbf{b}^t - \mathbf{b}^*)$ and $\mathbf{g} = (I - X^\top(XX^\top)^{-1}X)\boldsymbol{\epsilon}$ and analyzing the vector $X^\top\boldsymbol{\lambda}^t + \mathbf{g}$.

However, this is a bit cumbersome since $\mathbf{g}$ is not distributed as a spherical Gaussian, something we would like to be able to use in the subsequent proofs. To remedy this, we simply change notation to denote $\boldsymbol{\lambda}^t = (XX^\top)^{-1}(\mathbf{b}^t - \mathbf{b}^*) - (XX^\top)^{-1}X\boldsymbol{\epsilon}$ and $\mathbf{g} = \boldsymbol{\epsilon}$. This will not affect the results in the least since we have, as shown in the proof of Lemma 4, $\left\|(XX^\top)^{-1}X\boldsymbol{\epsilon}\right\|_2 \leq \sigma\sqrt{\frac{2d}{n}\log\frac{d}{\sigma}}$ because of which we can set n large enough so that $\left\|\boldsymbol{\lambda}^t\right\|_2 \leq \frac{\sigma}{100}$ still holds. Given this, we prove the following result:

Lemma 18. *Let $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ be a data matrix consisting of i.i.d. standard normal vectors i.e $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, and $\mathbf{g} \sim N(0, \sigma^2 \cdot I_{n\times n})$ be standard normal vector drawn independently of X. For any $\boldsymbol{\lambda} \in \mathbb{R}^d$ such that $\|\boldsymbol{\lambda}\|_2 \leq \frac{\sigma}{100}$, define $\mathbf{v} = X^\top\boldsymbol{\lambda} + \mathbf{g}$. For any $\tau > 0$, define the vector $\mathbf{z}$ such that $z_i = v_i$ if $|v_i| > \tau$ and $z_i = 0$ otherwise. Then, with probability at least $1-\delta$, for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ with norm at most $\frac{\sigma}{100}$, we have $\frac{1}{\lambda_n}\|X\mathbf{z}\|_2 \leq M(\tau)\|\boldsymbol{\lambda}\|_2 + 2.02\sigma\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}$, where $M(\tau) < \frac{0.808}{\sigma}\left(\tau + \frac{1}{\tau}\right)\exp\left(-\frac{\tau^2}{2.001\sigma^2}\right)$.*

Proof. We will first prove this result by first assuming that $\boldsymbol{\lambda}$ is a fixed d-dimensional vector with small norm and X and $\boldsymbol{\epsilon}$ are chosen independently of $\boldsymbol{\lambda}$. We will then generalize to all small norm vectors in $\mathbb{R}^d$ by taking a suitably fine ϵ-net over them. Let us denote the i^{th} row of X as X^i, and the entry at the j^{th} column in this row as X^i_j. Then $(X\mathbf{z})_i = \mathbf{z}^\top X^i = \sum_{j=1}^n X^i_j z_j$. Note that $v_j = \mathbf{x}_j^\top\boldsymbol{\lambda} + g_j$ and hence v_j and $v_{j'}$ are independent for $j \neq j'$. Because of this, $X^i_j z_j$ is also independent from $X^i_{j'} z_{j'}$.

We also note that $v_j | X^i_j \sim \mathcal{N}(X^i_j\lambda_i, \sigma^2 + \sum_{i'\neq i}\lambda^2_{i'})$. Let $\tilde{\sigma}^2 := \sigma^2 + \sum_{i'\neq i}\lambda^2_{i'}$. Note that $z_i = \mathbb{I}\{|v_i| > \tau\}\cdot v_i$. Using a simpler notation temporarily $x := X^i_j, z := z_j$ and $v := v_j$ lets us write

$$\mathbb{E}[xz] = \int_{\mathbb{R}\setminus[-\tau,\tau]}\int_{\mathbb{R}} xv\ p(x,v)\ dx\ dv.$$

Let $D_i := \left(I + \frac{\lambda_i^2}{\sigma^2}\right)^{1/2}$. Then for any fixed v, we have

$$\begin{aligned}
\int_{\mathbb{R}} xv\ p(x,v)\ dx &= \int_{\mathbb{R}} xv\ p(x)p(v|x)\ dx \\
&= \frac{1}{\tilde{\sigma}(\sqrt{2\pi})^2}\int_{\mathbb{R}} xv\ \exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{(v-x\lambda_i)^2}{2\tilde{\sigma}^2}\right)\ dx \\
&= \frac{vD_i^{-2}}{\tilde{\sigma}(\sqrt{2\pi})^2}\int_{\mathbb{R}} u\ \exp\left(-\frac{u^2}{2} + \frac{v^2}{\tilde{\sigma}^2} - \frac{2vuD_i^{-1}\lambda_i}{\tilde{\sigma}^2}\right)\ du \\
&= \frac{vD_i^{-2}\exp\left(-\frac{v^2}{2\tilde{\sigma}^2} + \frac{v^2D_i^{-2}\lambda_i^2}{2\tilde{\sigma}^4}\right)}{\tilde{\sigma}(\sqrt{2\pi})^2}\int_{\mathbb{R}} u\ \exp\left(-\frac{1}{2}\left(u - \frac{vD_i^{-1}\lambda_i}{\tilde{\sigma}^2}\right)^2\right)\ du \\
&= \frac{v^2D_i^{-3}\lambda_i\exp\left(-\frac{v^2}{2\tilde{\sigma}^2} + \frac{v^2D_i^{-2}\lambda_i^2}{2\tilde{\sigma}^4}\right)}{\tilde{\sigma}^3\sqrt{2\pi}} \\
&\leq \frac{v^2D_i^{-3}\lambda_i\exp\left(-\frac{v^2}{2.001\sigma^2}\right)}{1.001\sigma^3\sqrt{2\pi}},
\end{aligned}$$

where in the third step, we perform a change of variables $u = D_i x$ and in the last step, we use the fact that $\tilde{\sigma}^2 \leq \sigma^2 + \sigma^2/10000$ since $\|\boldsymbol{\lambda}\|_2 \leq \sigma/100$, as well as $\lambda_i^2 \leq \|\boldsymbol{\lambda}\|_2^2$. Plugging this into the expression for $\mathbb{E}[xz]$ and using elementary manipulations such as integration by parts gives us

$$\mathbb{E}\left[X^i_j z_j\right] = M(\tau)\lambda_i,\ \textit{i.e.},\ \ \mathbb{E}\left[\boldsymbol{\lambda}^T\mathbf{x}_j z_j\right] = M(\tau)\|\boldsymbol{\lambda}\|_2^2,$$

where $M(\tau) < 0.8\left(\frac{\tau}{\sigma} + \frac{\sigma}{\tau}\right)\exp\left(-\frac{\tau^2}{2.001\sigma^2}\right)$. This gives us $\mathbb{E}\left[\sum_{j=1}^n \boldsymbol{\lambda}^T\mathbf{x}_j z_j\right] = nM(\tau)\|\boldsymbol{\lambda}\|_2^2$. Moreover, for any j, $\boldsymbol{\lambda}^T\mathbf{x}_j$ is a $\|\boldsymbol{\lambda}\|_2$-subGaussian random variable and z_j is a 2σ-subGaussian random variable as

$\|\boldsymbol{\lambda}\|_2 \leq \sigma/100$. Hence, $\boldsymbol{\lambda}^T \mathbf{x}_j z_j$ is a sub-exponential random variable with sub-exponential norm $2\sigma\|\boldsymbol{\lambda}\|_2$. Using the Bernstein inequality for subexponential variables [21], then allows us to arrive at the following result, with probability at least $1-\delta$.

$$\sum_{j=1}^{n} \boldsymbol{\lambda}^T \mathbf{x}_j z_j \leq nM(\tau)\|\boldsymbol{\lambda}\|_2^2 + 2\sqrt{\sigma\|\boldsymbol{\lambda}\|_2}\sqrt{n\log\frac{2}{\delta}}.$$

Taking a union bound over an ϵ-net over all possible values of $\boldsymbol{\lambda}$ (i.e. which satisfy the norm bound), for $\epsilon = 1/100$ gives us, with probability at least $1-\delta$, for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\lambda}\|_2 \leq \frac{\sigma}{100}$,

$$\frac{1}{\lambda_n}\boldsymbol{\lambda}^T X\mathbf{z} \leq 1.01M(\tau)\,\|\boldsymbol{\lambda}\|_2^2 + 2.02\sqrt{\sigma\|\boldsymbol{\lambda}\|_2}\sqrt{\frac{d}{n}\log\frac{200}{\delta}}. \tag{8}$$

Now, again consider a fixed $\boldsymbol{\lambda}$ and a fixed *unit* vector $\mathbf{v} \in \mathbb{R}^d$ s.t. $\boldsymbol{\lambda}^T\mathbf{v} = 0$. In this case, $\mathbf{v}^T\mathbf{x}_j$ is independent of z_j. Hence, $\mathbb{E}\left[[\right] \mathbf{v}^T\mathbf{x}_j z_j] = 0$. Moreover, $\mathbf{v}^T\mathbf{x}_j z_j$ is a 2σ-subexponential random variable. Moreover, number of fixed $\boldsymbol{\lambda}$ and $\mathbf{v}$ in their ϵ-net is $\frac{1}{\epsilon}^d \cdot \frac{1}{\epsilon}^{d-1}$. Hence, using the subexponential Bernstein inequality and using union bound over all $\mathbf{v}$ and $\boldsymbol{\lambda}$, we get (w.p. $\geq 1-\delta$):

$$\max_{\mathbf{v},\boldsymbol{\lambda}} \frac{1}{\lambda_n}\mathbf{v}^T X\mathbf{z} \leq 2.02\sqrt{\sigma\frac{d}{n}\log\frac{200}{\delta}}. \tag{9}$$

Lemma now follows by using $\|X\mathbf{z}\|_2^2 = \frac{1}{\|\lambda\|_2^2}(\boldsymbol{\lambda}^T X\mathbf{z})^2 + \max_{\mathbf{v},\|\mathbf{v}\|_2=1,\mathbf{v}^T\boldsymbol{\lambda}=0}(\mathbf{v}^T X\mathbf{z})^2$ with (8) and (9).

This establishes the claimed result. □

Although Lemma 18 seems to close the issue of convergence of the iterates $\boldsymbol{\lambda}^t$, and hence the convergence of $\mathbf{w}^t$ and consistency, it is not so. The reason is twofold – firstly Lemma 18 works with a value based thresholding whereas CRR uses a cardinality based thresholding. Secondly, in order to establish a linear convergence rate for $\boldsymbol{\lambda}^t$, we need to show that the constant $M(\tau)$ is smaller than 98/100 so that we can ensure that $\left\|\boldsymbol{\lambda}^{t+1}\right\|_2 \leq \left(\frac{1}{100}+0.98\right)\left\|\boldsymbol{\lambda}^t\right\|_2 \leq 0.99\left\|\boldsymbol{\lambda}^t\right\|_2 + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right)$, thus ensuring a linear convergence for $\boldsymbol{\lambda}^t$, save negligible terms. We do both of these in the subsequent discussion.

We address both the above issues by showing that while thresholding the vector $X^\top\boldsymbol{\lambda}^t + \mathbf{g}$ (recall that for sake of notational convenience we are still omitting the subscript $\overline{S^*}$), the k^{th} top element in terms of magnitude will be large enough. Thus, thresholding at that value will recover the top k elements. If we are able to get a sample independent bound on the magnitude of this element then we can set τ to this in the analysis of Lemma 18 and be done. Of course, it will still have to be ensured that for this value of τ, we have $M(\tau) < 1$.

To simplify the discussion and calculations henceforth, we shall assume that $\sigma = 1$, $\delta = 1$, and $k = k^*$. We stress that all our analyses go through even for non-unit variance noise, projection parameters that differ from the true corruption sparsity (i.e. $k \neq k^*$), as well as can be readily modified to give high confidence bound. However, these assumptions greatly simplify our analyses.

We notice that the vector being thresholded has two components $X^\top\boldsymbol{\lambda}^t$ and $\mathbf{g}$. Whereas $\mathbf{g}$ has a nice characterization, being a standard Gaussian vector, there is very little we can say about the vector $X^\top\boldsymbol{\lambda}^t$ other than that the norm of the vector $\boldsymbol{\lambda}^t$ is small. This is because the vector $\boldsymbol{\lambda}^t$ is dependent on previous iterations and hence, dependent on X as well as $\mathbf{g}$. The way out of this is to show that the k^{th} largest element in $\mathbf{g}$ is reasonably large and $X^\top\boldsymbol{\lambda}^t$, on account of its small norm, cannot diminish it.

To proceed in this direction, we first recall the coarse convergence analysis. Letting $\alpha := \frac{k^*}{n}$ and making the assumptions stated above we know that $\|\boldsymbol{\lambda}^{T_0}\|_2 \leq \mathcal{C}(\alpha)$ where

$$\mathcal{C}(\alpha) = 2.001\sqrt{2\alpha}\sqrt{1+2e\sqrt{6\log\frac{e}{2\alpha}}}.$$

Note that $\lim_{\alpha\to 0}\mathcal{C}(\alpha)=0$, as well as that $\left\|X^\top\boldsymbol{\lambda}^{T_0}\right\|_2 \leq \mathcal{C}(\alpha)\cdot\sqrt{n}$. This bound gives us an idea about how much weight lies in the vector $X^\top\boldsymbol{\lambda}^t$ in the iterations $t>T_0$. Next we look at the other component $\mathbf{g}$. For any value $\eta>0$, the probability of a Gaussian variable exceeding that value in magnitude is given by $\sqrt{2}\cdot\text{erfc}(\eta/\sqrt{2})$, where erfc is the complimentary error function. By an application of Chernoff bounds, we can then conclude that in any ensemble of n such Gaussian variables, with probability at least $1-\exp(-\Omega(n))$ at least a $0.99\cdot\text{erfc}\left(\frac{\eta}{\sqrt{2}}\right)$ fraction (as well as at most a $1.01\cdot\text{erfc}\left(\frac{\eta}{\sqrt{2}}\right)$ fraction) of points will exceed the value η.

We also recall the quantity

$$M(\zeta)<0.8\left(\zeta+\frac{1}{\zeta}\right)\exp\left(-\frac{\zeta^2}{2.001}\right),$$

and notice that, in order for $M(\zeta)$ to get less than 98/100, ζ must be greater than 0.99. Now the previous estimate for bounds on Gaussian variables tells us that with probability at least $1-\exp(-\Omega(n))$, at least a $\beta=1/25$ fraction of values in the vector $\mathbf{g}$, which is a standard Gaussian (since we have assumed $\sigma=1$ for sake of simplicity) will exceed the value 1.98.

Let S_β denote the set of coordinates of $\mathbf{g}$ which exceed the value 1.98. Let us call a coordinate $i\in S_\beta$ *corrupted* if $\left|(X^\top\boldsymbol{\lambda}^{T_0})_i\right|\geq 0.98$. Now we notice that if this happens for $(\beta-\alpha)\cdot n$ points in the set S_β, then $\left\|X^\top\boldsymbol{\lambda}^{T_0}\right\|_2\geq 0.98\sqrt{(\beta-\alpha)n}$. Thus, we set $\mathcal{C}(\alpha)\cdot\sqrt{n}<0.98\sqrt{(\beta-\alpha)n}=0.98\sqrt{(0.04-\alpha)n}$ to prevent this from happening. We note that for all values of $\alpha<\frac{1}{10000}$ this is true. This ensures that at least $k^*=\alpha\cdot n$ points in the set S are of magnitude at least 1 and thus we can set $\tau=1$ in Lemma 18 which then finishes the proof since $M(1)<0.98$. □

B Supplementary Material for Consistent Robust Time Series Estimation

B.1 Main Result

Theorem 9. *Let* $\mathbf{y}$ *be generated using* $AR(d)$ *process with* k^* *additive outliers (see* (3)*). Also, let* $k^*\leq k\leq C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$ *(for some universal constant* $C>0$*). Then, with probability at least* $1-\delta$, CRTSE*, after* $\mathcal{O}(\log(\|\mathbf{b}^*\|_2/n)+\log(n/(\sigma\cdot d)))$ *steps, ensures that* $\|\mathbf{w}^t-\mathbf{w}^*\|_2\leq\mathcal{O}\left(\sigma\mathcal{M}_{\mathbf{w}^*}/\mathfrak{m}_{\mathbf{w}^*}\sqrt{d\log n/n\log(d/\delta)}\right)$.

Proof. Putting together the Lemma 11 and the equation (7) establishes that

$$\left\|\boldsymbol{\lambda}^{t+1}\right\|_2\leq 0.51\left\|\boldsymbol{\lambda}^t\right\|_2+\mathcal{O}\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{d}{\delta}}\right),$$

which ensures a linear convergence of the terms $\left\|\boldsymbol{\lambda}^t\right\|_2$ to a value $\mathcal{O}\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{d}{\delta}}\right)$. Applying the equation (6), then finishes off the result. □

B.2 Back ground on Time Series

AR (d) process is defined as

$$x_t \;=\; x_{t-1}\mathbf{w}_1^*+\cdots+x_{t-d}\mathbf{w}_d^*+\boldsymbol{\epsilon}_t \text{ where } \boldsymbol{\epsilon}_t\sim\mathcal{N}(0,\sigma^2). \tag{10}$$

Note that $x_t \sim \mathcal{N}(0, \Gamma(0))$, where $\Gamma(h) = \mathbb{E}[x_t x_{t+h}]$ is the auto-covariance function of the time series. Then we have

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_0 & \cdots & x_{-d+1} \\ \vdots & & \vdots \\ x_{n-1} & \cdots & x_{n-d} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{w}_1^* \\ \vdots \\ \mathbf{w}_d^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{bmatrix}$$

$$\mathbf{y}^* = \overline{X}^\top \mathbf{w}^* + \boldsymbol{\epsilon}. \tag{11}$$

The spectral density of this AR(d) process can be given as

$$\rho_{\mathbf{w}^*}(\omega) = \frac{\sigma^2}{\left(1 - \sum_{k=1}^d \mathbf{w}_k^* e^{ik\omega}\right)\left(1 - \sum_{k=1}^d \mathbf{w}_k^* e^{-ik\omega}\right)}, \text{ for } \omega \in [0, 2\pi]. \tag{12}$$

Observe that any column vector of the matrix $\overline{X}$ is distributed as $\overline{X}_i \sim \mathcal{N}(0, C_{\overline{X}})$, where

$$C_{\overline{X}} = \begin{bmatrix} \Gamma(0) & \Gamma(1) & \cdots & \Gamma(d-1) \\ \Gamma(1) & \Gamma(0) & \cdots & \Gamma(d-2) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma(d-1) & \Gamma(d-2) & \cdots & \Gamma(0) \end{bmatrix}.$$

Since $C_{\overline{X}}$ is a block-Toeplitz matrix, we have

$$\mathfrak{m}_{\mathbf{w}^*} := \inf_{\omega \in [0,2\pi]} \rho_{\mathbf{w}^*}(\omega) \leq \Lambda_{\min}[C_{\overline{X}}] \leq \Lambda_{\max}[C_{\overline{X}}] \leq \sup_{\omega \in [0,2\pi]} \rho_{\mathbf{w}^*}(\omega) =: \mathcal{M}_{\mathbf{w}^*}. \tag{13}$$

The columns of $\overline{X}$ can be viewed as a d-variate of VAR(1) process as follows

$$\begin{bmatrix} x_i \\ x_{i-1} \\ \vdots \\ x_{i-(d-1)} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^* & \mathbf{w}_2^* & \cdots & \mathbf{w}_{d-1}^* & \mathbf{w}_d^* \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_{i-1} \\ x_{i-2} \\ \vdots \\ x_{i-d} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\widehat{X}_i = W\widehat{X}_{i-1} + \mathcal{E}_i, \text{ for } i = 1, \ldots, n. \tag{14}$$

By letting

$$\mathbf{u}^* = \begin{bmatrix} \widehat{X}_1 \\ \vdots \\ \widehat{X}_n \end{bmatrix} \in \mathbb{R}^{nd}, \ \mathcal{U} = \begin{bmatrix} \widehat{X}_0 \\ \vdots \\ \widehat{X}_{n-1} \end{bmatrix} \in \mathbb{R}^{nd}, \text{ and } \mathcal{E} = \begin{bmatrix} \mathcal{E}_1 \\ \vdots \\ \mathcal{E}_n \end{bmatrix} \in \mathbb{R}^{nd}$$

the above VAR(1) process can be compactly written as follows

$$\mathbf{u}^* = W\mathcal{U} + \mathcal{E}.$$

Then the spectral density of the above VAR(1) process is given by

$$\rho_W(\omega) = \left(I - We^{-i\omega}\right)^{-1} \Sigma_{\boldsymbol{\epsilon}} \left[\left(I - We^{-i\omega}\right)^{-1}\right]^*, \text{ for } \omega \in [0, 2\pi],$$

where

$$\Sigma_{\boldsymbol{\epsilon}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The covariance matrix of vector $\mathcal{U}$ is given by

$$C_{\mathcal{U}} \;=\; \mathbb{E}\left[\mathcal{U}\mathcal{U}^\top\right] \;=\; \begin{bmatrix} \mathbb{E}\left[\widehat{X}_0\widehat{X}_0^\top\right] & \mathbb{E}\left[\widehat{X}_0\widehat{X}_1^\top\right] & \cdots & \mathbb{E}\left[\widehat{X}_0\widehat{X}_{n-1}^\top\right] \\ \mathbb{E}\left[\widehat{X}_1\widehat{X}_0^\top\right] & \mathbb{E}\left[\widehat{X}_1\widehat{X}_1^\top\right] & \cdots & \mathbb{E}\left[\widehat{X}_1\widehat{X}_{n-1}^\top\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\widehat{X}_{n-1}\widehat{X}_0^\top\right] & \mathbb{E}\left[\widehat{X}_{n-1}\widehat{X}_1^\top\right] & \cdots & \mathbb{E}\left[\widehat{X}_{n-1}\widehat{X}_{n-1}^\top\right] \end{bmatrix}.$$

Since $C_{\mathcal{U}}$ is a block-Toeplitz matrix, we have

$$\Lambda_{\max}\left[C_{\mathcal{U}}\right] \le \sup_{\omega\in[0,2\pi]} \rho_W(\omega) = \frac{\sigma^2}{\inf_{\omega\in[0,2\pi]} \Lambda_{\min}\left[\left(I - W^\top e^{i\omega}\right)\left(I - We^{-i\omega}\right)\right]} =: \mathcal{M}_W. \tag{15}$$

Consider a vector $\mathbf{q} = \overline{X}^\top \mathbf{a} \in \mathbb{R}^n$ for any $\mathbf{a} \in S^{d-1}$. Since each element $\overline{X}_i^\top \mathbf{a} \sim \mathcal{N}\left(0, \mathbf{a}^\top C_{\overline{X}}\mathbf{a}\right)$, it follows that $\mathbf{q} \sim \mathcal{N}\left(0, Q_{\mathbf{a}}\right)$ where $Q_{\mathbf{a}} = \left(I_n \otimes \mathbf{a}^\top\right) C_{\mathcal{U}} \left(I_n \otimes \mathbf{a}\right)$. From this we can note that

$$\begin{aligned} \operatorname{trace}\left(Q_{\mathbf{a}}\right) \;&=\; n\mathbf{a}^\top C_{\overline{X}}\mathbf{a} \;\le\; n\Lambda_{\max}\left[C_{\overline{X}}\right] \;\le\; n\mathcal{M}_{\mathbf{w}^*} && (16)\\ \left\|Q_{\mathbf{a}}\right\|_2 \;&\le\; \left\|\mathbf{a}\right\|_2^2 \Lambda_{\max}\left[C_{\mathcal{U}}\right] \;\le\; \mathcal{M}_W && (17)\\ \left\|Q_{\mathbf{a}}\right\|_{\mathrm{F}} \;&=\; \sqrt{\operatorname{trace}\left(Q_{\mathbf{a}}Q_{\mathbf{a}}\right)} \;\le\; \sqrt{\left\|Q_{\mathbf{a}}\right\|_2 \operatorname{trace}\left(Q_{\mathbf{a}}\right)} \;\le\; \sqrt{n\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}}. && (18) \end{aligned}$$

Additive Corruptions: Now consider the following additive corruption mechanism (at most k^* data points):

$$\mathbf{y}_i \;=\; \mathbf{y}_i^* + \mathbf{e}_i^* \;=\; x_i + \mathbf{e}_i^* \text{ for } i = 1, \ldots, n.$$

Since we observe the corrupted time series data $(y_{-d+1}, \ldots, y_0, y_1, \ldots, y_n)$, we have

$$\begin{bmatrix} y_0 & y_{-1} & \cdots & y_{-d+1} \\ y_1 & y_0 & \cdots & y_{-d+2} \\ y_2 & y_1 & \cdots & y_{-d+3} \\ \vdots & \vdots & & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-d} \end{bmatrix} = \begin{bmatrix} x_0 & x_{-1} & \cdots & x_{-d+1} \\ x_1 & x_0 & \cdots & x_{-d+2} \\ x_2 & x_1 & \cdots & x_{-d+3} \\ \vdots & \vdots & & \vdots \\ x_{n-1} & x_{n-2} & \cdots & x_{n-d} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \mathbf{e}_1^* & 0 & \cdots & 0 \\ \mathbf{e}_2^* & \mathbf{e}_1^* & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \mathbf{e}_{n-1}^* & \mathbf{e}_{n-2}^* & \cdots & \mathbf{e}_{n-d}^* \end{bmatrix}$$

$$X^\top \;=\; \overline{X}^\top + E^\top \tag{19}$$

Thus the observed time series can be modeled as follows

$$\begin{aligned} \mathbf{y} \;&=\; \mathbf{y}^* + \mathbf{e}^* \\ &=\; \overline{X}^\top \mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{e}^* \\ &=\; (X^\top - E^\top)\mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{e}^* \\ &=\; X^\top \mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{b}^*_{\mathbf{e}^*,\mathbf{w}^*}, \end{aligned} \tag{20}$$

where $\mathbf{e}^* = (\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*)^\top$ is k^*-sparse, and $\mathbf{b}^*_{\mathbf{e}^*,\mathbf{w}^*} = \mathbf{e}^* - E^\top\mathbf{w}^*$ is k^*-block-sparse with block size of $d+1$ (since $E\mathbf{w}^*$ is k^*-block-sparse with block size of d).

B.3 Singular values of $\overline{X}$

Lemma 19. *Let $\overline{X}$ be a matrix whose columns are sampled from a stationary and stable VAR(1) process given by (14) i.e. $\overline{X}_i \sim \mathcal{N}\left(0, C_{\overline{X}}\right)$. Then for any $\epsilon > 0$, with probability at least $1-\delta$, $\overline{X}$ satisfies*

$$\lambda_{\max}\left(\overline{X}\,\overline{X}^\top\right) \;\le\; n\mathcal{M}_{\mathbf{w}^*} + (1-2\epsilon)^{-1}\left\{\sqrt{n\alpha_1(d,\delta,\epsilon)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + \alpha_1(d,\delta,\epsilon)\mathcal{M}_W\right\}$$

$$\lambda_{\min}\left(\overline{XX}^\top\right) \;\geq\; n\mathfrak{m}_{\mathbf{w}^*} - (1-2\epsilon)^{-1}\left\{\sqrt{n\alpha_1(d,\delta,\epsilon)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + \alpha_1(d,\delta,\epsilon)\mathcal{M}_W\right\},$$

where $\alpha_1(d,\delta,\epsilon) = c\log\frac{2}{\delta} + cd\log\frac{3}{\epsilon}$ *for some universal constant* c.

Proof. Using the results from [4, 14], we first show that with high probability,

$$\left\|\overline{XX}^\top - nC_{\overline{X}}\right\|_2 \;\leq\; \epsilon_1$$

for some $\epsilon > 0$. Doing so will automatically establish the following result

$$n\Lambda_{\min}\left[C_{\overline{X}}\right] - \epsilon_1 \;\leq\; \lambda_{\min}\left(\overline{XX}^\top\right) \;\leq\; \lambda_{\max}\left(\overline{XX}^\top\right) \;\leq\; n\Lambda_{\max}\left[C_{\overline{X}}\right] + \epsilon_1.$$

Let $C^{d-1}(\epsilon) \subset S^{d-1}$ be an ϵ-cover of S^{d-1} ([21], see Definition 5.1). Standard constructions ([21], see Lemma 5.2) guarantee such a cover of size at most $\left(1+\frac{2}{\epsilon}\right)^d \leq \left(\frac{3}{\epsilon}\right)^d$. Further by Lemma 5.4 from [21], we have

$$\left\|\overline{XX}^\top - nC_{\overline{X}}\right\|_2 \;\leq\; (1-2\epsilon)^{-1}\sup_{\mathbf{u}\in C^{d-1}(\epsilon)}\left|\mathbf{u}^\top\left(\overline{XX}^\top - nC_{\overline{X}}\right)\mathbf{u}\right|.$$

By following the analysis given in [4, 14], we can provide a high probability bound on $\left|\mathbf{u}^\top\left(\overline{XX}^\top - nC_{\overline{X}}\right)\mathbf{u}\right|$. For any $\mathbf{u} \in S^{d-1}$, let $\mathbf{q} = \overline{X}^\top\mathbf{u} \sim \mathcal{N}(0, Q_{\mathbf{u}})$ where $Q_{\mathbf{u}} = \left(I_n \otimes \mathbf{u}^\top\right) C_{\mathcal{U}}\left(I_n \otimes \mathbf{u}\right)$. Note that $\mathbf{u}^\top\overline{XX}^\top\mathbf{u} = \mathbf{q}^\top\mathbf{q} = \mathbf{z}^\top Q_{\mathbf{u}}\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, I_n)$. Also, $\mathbf{u}^\top nC_{\overline{X}}\mathbf{u} = \mathbb{E}\left[\mathbf{z}^\top Q_{\mathbf{u}}\mathbf{z}\right]$. So, by the Hanson-Wright inequality of [18], with $\|\mathbf{z}_i\|_{\psi_2} \leq 1$ since $\mathbf{z}_i \sim \mathcal{N}(0,1)$, we get

$$\begin{aligned}\mathbb{P}\left[\left|\mathbf{u}^\top\left(\overline{XX}^\top - nC_{\overline{X}}\right)\mathbf{u}\right| > \lambda\right] \;&=\; \mathbb{P}\left[\left|\mathbf{z}^\top Q_{\mathbf{u}}\mathbf{z} - \mathbb{E}\left[\mathbf{z}^\top Q_{\mathbf{u}}\mathbf{z}\right]\right| > \lambda\right] \\ &\leq\; 2\exp\left(-\frac{1}{c}\min\left\{\frac{\lambda^2}{\|Q_{\mathbf{u}}\|_F^2}, \frac{\lambda}{\|Q_{\mathbf{u}}\|_2}\right\}\right).\end{aligned}$$

Setting $\lambda = \sqrt{\alpha_1(d,\delta,\epsilon)}\,\|Q_{\mathbf{u}}\|_F + \alpha_1(d,\delta,\epsilon)\,\|Q_{\mathbf{u}}\|_2$, and taking a union bound over all $C^{d-1}(\epsilon)$, we get

$$\begin{aligned}&\mathbb{P}\left[\sup_{\mathbf{u}\in C^{d-1}(\epsilon)}\left|\mathbf{u}^\top\left(\overline{XX}^\top - nC_{\overline{X}}\right)\mathbf{u}\right| > \sqrt{\alpha_1(d,\delta,\epsilon)}\,\|Q_{\mathbf{u}}\|_F + \alpha_1(d,\delta,\epsilon)\,\|Q_{\mathbf{u}}\|_2\right] \\ \leq\; 2\left(\frac{3}{\epsilon}\right)^d &\exp\left(-\frac{1}{c}\min\left\{\frac{\lambda^2}{\|Q_{\mathbf{u}}\|_F^2}, \frac{\lambda}{\|Q_{\mathbf{u}}\|_2}\right\}\right) \;\leq\; \delta.\end{aligned}$$

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

$$\left\|\overline{XX}^\top - nC_{\overline{X}}\right\|_2 \;\leq\; (1-2\epsilon)^{-1}\left\{\sqrt{\alpha_1(d,\delta,\epsilon)}\,\|Q_{\mathbf{u}}\|_F + \alpha_1(d,\delta,\epsilon)\,\|Q_{\mathbf{u}}\|_2\right\},$$

which (along with the bounds given in (16),(17), and (18)) gives us the claimed bounds on the singular values of $\overline{XX}^\top$. □

B.4 Restricted Singular values of $\overline{X}$

Lemma 20. *Let* $\overline{X}$ *be a matrix whose columns are sampled from a stationary and stable VAR*(1) *process given by* (14) *i.e.* $\overline{X}_i \sim \mathcal{N}(0, C_{\overline{X}})$. *Then for any* $k \leq \frac{n}{d}$, *with probability at least* $1-\delta$, *the matrix* $\overline{X}$ *satisfies the SGSC and SGSS properties with constants*

$$\Lambda_k \;\leq\; k\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}\frac{1}{c}\log\frac{en}{kd}} + \mathcal{M}_W\log\frac{en}{kd}\right\}$$

$$
\begin{aligned}
&+ \mathcal{O}\left(\sqrt{kd\alpha_2(d,\delta)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}}\right) + \mathcal{O}\left(\alpha_2(d,\delta)\mathcal{M}_W\right) \\
\lambda_k \;&\geq\; n\mathfrak{m}_{\mathbf{w}^*} - \left(\frac{n}{d} - k\right)\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}\frac{1}{c}\log\frac{en}{n-kd}} + \log\frac{en}{n-kd}\mathcal{M}_W\right\} \\
&- \Omega\left(\left(1+\sqrt{\frac{n-kd}{n}}\right)\sqrt{n\alpha_2(d,\delta)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}}\right) - \Omega\left(\alpha_2(d,\delta)\mathcal{M}_W\right),
\end{aligned}
$$

where $\alpha_2(d,\delta) = \log\frac{1}{\delta} + d$ *and* c *is some universal constant.*

Proof. One can easily observe that considering the columns-restricted matrix $\overline{X}_S$ wouldn't impact the analysis of Lemma 19. Thus for any fixed $S \in \mathcal{S}_k^{\mathcal{G}}$, Lemma 19 guarantees the following bound (since $|S| = kd$)

$$
\lambda_{\max}\left(\overline{X}_S\overline{X}_S^\top\right) \;\leq\; kd\mathcal{M}_{\mathbf{w}^*} + (1-2\epsilon)^{-1}\left\{\sqrt{kd\alpha_1(d,\delta,\epsilon)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + \alpha_1(d,\delta,\epsilon)\mathcal{M}_W\right\}.
$$

Taking a union bound over $\mathcal{S}_k^{\mathcal{G}}$ and noting that $\left|\mathcal{S}_k^{\mathcal{G}}\right| \leq \left(\frac{en}{kd}\right)^k$, gives us with probability at least $1-\delta$

$$
\Lambda_k \;\leq\; kd\mathcal{M}_{\mathbf{w}^*} + (1-2\epsilon)^{-1}M,
$$

where

$$
\begin{aligned}
M \;&=\; \sqrt{\left(\alpha_1(d,\delta,\epsilon) + ck\log\frac{en}{kd}\right)kd\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + \left(\alpha_1(d,\delta,\epsilon) + ck\log\frac{en}{kd}\right)\mathcal{M}_W \\
&\leq\; \sqrt{\alpha_1(d,\delta,\epsilon)kd\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + k\sqrt{cd\log\frac{en}{kd}\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + \left(\alpha_1(d,\delta,\epsilon) + ck\log\frac{en}{kd}\right)\mathcal{M}_W.
\end{aligned}
$$

If $c < 1$ (which can be ensured by scaling), by setting $\epsilon = \frac{1}{2}(1-c)$ and noting that $\Theta\left(\frac{1}{c}\alpha_1\left(d,\delta,\frac{1-c}{2}\right)\right) = \Theta\left(\log\frac{1}{\delta} + d\right)$, we get

$$
\begin{aligned}
\Lambda_k \;&\leq\; k\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}\frac{1}{c}\log\frac{en}{kd}} + \mathcal{M}_W\log\frac{en}{kd}\right\} \\
&+ \mathcal{O}\left(\sqrt{kd\left(\log\frac{1}{\delta} + d\right)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}}\right) + \mathcal{O}\left(\left(\log\frac{1}{\delta} + d\right)\mathcal{M}_W\right).
\end{aligned}
$$

For the second bound, we use the equality

$$
\overline{X}_S\overline{X}_S^\top = \overline{XX}^\top - \overline{X}_{\overline{S}}\overline{X}_{\overline{S}}^\top,
$$

which provides the following bound for λ_k,

$$
\lambda_k \;\geq\; \lambda_{\min}\left(\overline{XX}^\top\right) - \max_{T\in\mathcal{S}_{\frac{n}{d}-k}^{\mathcal{G}}}\lambda_{\max}\left(\overline{X}_T\overline{X}_T^\top\right) \;=\; \lambda_{\min}\left(\overline{XX}^\top\right) - \Lambda_{\frac{n}{d}-k}.
$$

Using Lemma 19 to bound the first quantity and the first part of this theorem to bound the second quantity gives us, with probability at least $1-\delta$,

$$
\begin{aligned}
\lambda_k \;&\geq\; n\mathfrak{m}_{\mathbf{w}^*} - \left(\frac{n}{d} - k\right)\left\{d\mathcal{M}_{\mathbf{w}^*} + (1-2\epsilon)^{-1}\left(\sqrt{cd\log\frac{en}{n-kd}\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + c\log\frac{en}{n-kd}\mathcal{M}_W\right)\right\} \\
&- (1-2\epsilon)^{-1}\left\{\left(1+\sqrt{\frac{n-kd}{n}}\right)\sqrt{n\alpha_1(d,\delta,\epsilon)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}} + 2\alpha_1(d,\delta,\epsilon)\mathcal{M}_W\right\}.
\end{aligned}
$$

By setting $\epsilon = \frac{1}{2}(1-c)$ we get the following bound

$$\begin{aligned}\lambda_k \;\geq\; & n\mathfrak{m}_{\mathbf{w}^*} - \left(\frac{n}{d} - k\right)\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}\frac{1}{c}\log\frac{en}{n-kd}} + \log\frac{en}{n-kd}\mathcal{M}_W\right\} \\ & - \Omega\left(\left(1+\sqrt{\frac{n-kd}{n}}\right)\sqrt{n\left(\log\frac{1}{\delta}+d\right)\mathcal{M}_W\mathcal{M}_{\mathbf{w}^*}}\right) - \Omega\left(\left(\log\frac{1}{\delta}+d\right)\mathcal{M}_W\right).\end{aligned}$$

□

Remark 21. *Note that $\mathcal{M}_{\mathbf{w}^*}, \mathcal{M}_W$ and $\mathfrak{m}_{\mathbf{w}^*}$ will depend only on the actual model parameter vector $\mathbf{w}^*$ and σ (not on the realized data). Moreover $\mathcal{M}_{\mathbf{w}^*}$ and $\mathcal{M}_W$ are closely related. For example, for $AR(1)$ time-series with $0 < \mathbf{w}_1^* < 1$, we have $\mathcal{M}_{\mathbf{w}^*} = \mathcal{M}_W = \frac{\sigma^2}{(1-\mathbf{w}_1^*)^2}$ and $\mathfrak{m}_{\mathbf{w}^*} = \frac{\sigma^2}{(1+\mathbf{w}_1^*)^2}$. Then for sufficiently large enough n so that $\sqrt{n} \ll n$, the restricted singular value bounds of $\overline{X}$ from Lemma 20 can be simplified as follows*

$$\begin{aligned}\Lambda_k \;&\leq\; \mathcal{O}\left(k\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_{\mathbf{w}^*}\mathcal{M}_W\log\frac{en}{\delta kd}} + \mathcal{M}_W\log\frac{en}{\delta kd}\right\}\right) \text{ and} \\ \lambda_{\frac{n}{d}} \;&\geq\; \Omega\left(n\mathfrak{m}_{\mathbf{w}^*}\right).\end{aligned}$$

B.5 Restricted Singular values of X

Theorem 22 (SGSS/SGSC in AR(d) with AO model)**.** *Let X be the matrix given in* (3) *(additive corrupted $AR(d)$ model setting). Then for any $k \leq \frac{n}{d}$ and sufficiently large enough n, with probability at least $1-\delta$, the matrix X satisfies the SGSC and SGSS properties with constants*

$$\begin{aligned}\Lambda_k \;&\leq\; \mathcal{O}\left(k\left\{d\log n\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_{\mathbf{w}^*}\mathcal{M}_W\log\frac{en}{\delta kd}} + \mathcal{M}_W\log\frac{en}{\delta kd}\right\}\right) \\ \lambda_{\frac{n}{d}} \;&\geq\; \Omega\left(n\mathfrak{m}_{\mathbf{w}^*}\right).\end{aligned}$$

Proof. Recall that the matrix X can be decomposed as follows

$$X \;=\; \overline{X} + E.$$

Since for any $\mathbf{v} \in S^{n-1}$, $\|E\mathbf{v}\|_2^2 = \sum_{i=1}^d \left\langle (E^\top)_i, \mathbf{v}\right\rangle^2 \leq \sum_{i=1}^d \left\|(E^\top)_i\right\|_2^2 \|\mathbf{v}\|_2^2 \leq d\,\|\mathbf{e}^*\|_2^2$, we get $\|E\|_2 \leq \sqrt{d}\,\|\mathbf{e}^*\|_2$. By using the inequality $\left\|X_S - \overline{X}_S\right\|_2 \leq \|E_S\|_2 \leq \|E\|_2$ we get

$$\Lambda_{\min}\left[\overline{X}_S\right] - \|E\|_2 \;\leq\; \Lambda_{\min}\left[X_S\right] \;\leq\; \Lambda_{\max}\left[X_S\right] \;\leq\; \Lambda_{\max}\left[\overline{X}_S\right] + \|E\|_2.$$

Since $\mathbf{e}^*$ is k^*-sparse and $\mathbf{e}_i^* \leq \widehat{\sigma} = \mathcal{O}(\sqrt{\log n}\sigma)$, we have $\|\mathbf{e}^*\|_2 \leq \mathcal{O}(\sqrt{k^*\log n}\sigma)$. Thus from Lemma 20 and Remark 21, for sufficiently large enough n (with probability at least $1-\delta$) we get

$$\begin{aligned}\sqrt{\Lambda_k} \;&\leq\; \mathcal{O}\left(\sqrt{k\left\{d\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_{\mathbf{w}^*}\mathcal{M}_W\log\frac{en}{\delta kd}} + \mathcal{M}_W\log\frac{en}{\delta kd}\right\}}\right) + \mathcal{O}(\sqrt{k^*d\log n}\sigma) \\ &\leq\; \mathcal{O}\left(\sqrt{k\left\{d\log n\mathcal{M}_{\mathbf{w}^*} + \sqrt{d\mathcal{M}_{\mathbf{w}^*}\mathcal{M}_W\log\frac{en}{\delta kd}} + \mathcal{M}_W\log\frac{en}{\delta kd}\right\}}\right) \\ \sqrt{\lambda_{\frac{n}{d}}} \;&\geq\; \Omega\left(\sqrt{n\mathfrak{m}_{\mathbf{w}^*}}\right) - \Omega(\sqrt{k^*d\log n}\sigma) \;\geq\; \Omega\left(\sqrt{n\mathfrak{m}_{\mathbf{w}^*}}\right),\end{aligned}$$

which completes the proof. □

Remark 23. *Using Theorem 22, we can bound* $\frac{\sqrt{\Lambda_{k+k^*}}}{\lambda_{\frac{n}{d}}}$ *(which is required for the coarse convergence analysis of* CRTSE*) as follows (with probability at least* $1-\delta$*, and sufficiently large enough* n*)*

$$\begin{aligned}\frac{\sqrt{\Lambda_{k+k^*}}}{\lambda_{\frac{n}{d}}} &\leq \mathcal{O}\left(\frac{1}{n\mathfrak{m}_{\mathbf{w}^*}}\sqrt{k\left\{d\log n\mathcal{M}_{\mathbf{w}^*}+\sqrt{d\mathcal{M}_{\mathbf{w}^*}\mathcal{M}_W\log\frac{en}{\delta kd}}+\mathcal{M}_W\log\frac{en}{\delta kd}\right\}}\right)\\ &= \frac{f(\mathbf{w}^*,\sigma)\sqrt{\log n}}{n}\sqrt{(k+k^*)\left(d+2e\sqrt{6d\log\frac{en}{\delta(k+k^*)d}}\right)},\end{aligned}$$

for some positive function $f(\mathbf{w}^*,\sigma)$ *(suppressing* $\mathcal{M}_{\mathbf{w}^*},\mathcal{M}_W$ *and* $\mathfrak{m}_{\mathbf{w}^*}$*).*

From Theorem 22, it can also be observed that, if $k \leq C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$ *(for some universal constant* $C>0$*), then with probability at least* $1-\delta$*, we get* $\frac{\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}} \leq \frac{\Lambda_{2k}}{\lambda_{\frac{n}{d}}} \leq \frac{1}{4}$.

B.6 Bound on $\|X\boldsymbol{\epsilon}\|_2$

Lemma 24. *Let* X *be the matrix given in* (3) *(additive corrupted* $AR(d)$ *model setting). Then with probability at least* $1-\delta$*,*

$$\|X\boldsymbol{\epsilon}\|_2 \leq 2\sigma\sqrt{n\sqrt{\log n}c'd\log\frac{2d}{\delta}}.$$

for some constant $c'>0$.

Proof. We first bound the absolute value of $(\overline{X}\boldsymbol{\epsilon})_i = \sum_{j=1}^n \boldsymbol{\epsilon}_j x_{j-i}$ for $i=1,\ldots,d$. Let $z_j := \boldsymbol{\epsilon}_j x_{j-i}$. Since $\mathbb{E}[z_j|\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_{j-1}]=0$, $\{z_j : j\in[n]\}$ is a martingale difference sequence w.r.t $\{\boldsymbol{\epsilon}_j : j\in[n]\}$. Also note that for any j, $(z_j|\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_{j-1}) \sim \mathcal{N}(0,x_{j-i}^2\sigma^2)$. Then using the tail bounds on Gaussian random variables we have

$$\mathbb{P}\left[|z_j|>t|\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_{j-1}\right] \leq \sqrt{\frac{2}{\pi}}\frac{1}{|x_{j-i}|\,\sigma}\exp\left(\frac{-t^2}{2x_{j-i}^2\sigma^2}\right) \leq \sqrt{\frac{2}{\pi}}\frac{1}{c\sqrt{\log n}\sigma^2}\exp\left(\frac{-t^2}{2c^2\log n\sigma^4}\right),$$

since $\sup_{i\in[n]}|x_i| \leq \mathcal{O}\left(\sqrt{\log n}\sigma\right)$ with high probability. Then by using Theorem 2 from [19], we get

$$\mathbb{P}\left[\sum_{j=1}^n z_j > n\epsilon\right] \leq \exp\left(\frac{-\frac{1}{2c^2\log n\sigma^4}n\epsilon^2}{28\sqrt{\frac{2}{\pi}}\frac{1}{c\sqrt{\log n}\sigma^2}}\right) = \exp\left(\frac{-n\epsilon^2}{c'\sqrt{\log n}\sigma^2}\right),$$

for some constant c'. Similarly we also have

$$\mathbb{P}\left[\sum_{j=1}^n z_j < -n\epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{c'\sqrt{\log n}\sigma^2}\right).$$

Then by using the union bound we get (for any $\delta>0$)

$$\mathbb{P}\left[\left|\sum_{j=1}^n z_j\right| > n\epsilon\right] \leq 2\exp\left(\frac{-n\epsilon^2}{c'\sqrt{\log n}\sigma^2}\right) = \delta.$$

That is with probability at least $1-\delta$ we have

$$\left|(\overline{X}\boldsymbol{\epsilon})_i\right| \leq n\epsilon \leq \sigma\sqrt{n\sqrt{\log n}c'\log\frac{2}{\delta}}.$$

Taking a union bound gives us, with the same confidence,

$$\left\|\overline{X}\boldsymbol{\epsilon}\right\|_2^2 \leq \sigma^2 n\sqrt{\log n}c' d\log\frac{2d}{\delta}.$$

Now we bound the absolute value of $(E\boldsymbol{\epsilon})_i = \sum_{j=1}^n \boldsymbol{\epsilon}_j E_{i,j}$ for $i = 1,\ldots,d$. Let $z_j := \boldsymbol{\epsilon}_j E_{i,j}$. Note that for any j, $(z_j|\boldsymbol{\epsilon}_1,\ldots,\boldsymbol{\epsilon}_{j-1}) \sim \mathcal{N}(0, E_{i,j}^2\sigma^2)$ and $\sup_{i,j\in[n]}|E_{i,j}| \leq \widehat{\sigma} = \mathcal{O}\left(\sqrt{\log n}\sigma\right)$. Then by following the similar analysis as above, we have with probability at least $1-\delta$,

$$\|E\boldsymbol{\epsilon}\|_2^2 \leq \sigma^2 n\sqrt{\log n}c' d\log\frac{2d}{\delta}.$$

Then using the triangular inequality, with probability at least $1-\delta$,

$$\|X\boldsymbol{\epsilon}\|_2 \leq \left\|\overline{X}\boldsymbol{\epsilon}\right\|_2 + \|E\boldsymbol{\epsilon}\|_2 \leq 2\sigma\sqrt{n\sqrt{\log n}c' d\log\frac{2d}{\delta}}.$$

□

B.7 Coarse Convergence Analysis

Theorem 10. *For any data matrix X that satisfies the SGSC and SGSS properties such that $\frac{4\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}} < 1$,* CRTSE*, when executed with a parameter $k \geq k^*$, ensures that after $T_0 = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{\sqrt{n}}\right)$ steps, $\left\|\mathbf{b}^{T_0} - \mathbf{b}^*\right\|_2 \leq 5e_0$, where $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)d\log\frac{n}{\delta(k+k^*)d}}\right)$ for standard Gaussian $AR(d)$ process. If k is sufficiently small i.e. $k^* \leq k \leq C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$ (for some universal constant $C > 0$) and n is sufficiently large enough, then with probability at least $1-\delta$, we have $\frac{4\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}} < 1$.*

Proof. We start with the update step in CRTSE, and use the fact that $\mathbf{y} = X^\top\mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{b}^*$ to rewrite the update as

$$\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k^{\mathcal{G}}(P_X\mathbf{b}^t + (I-P_X)(X^\top\mathbf{w}^* + \boldsymbol{\epsilon} + \mathbf{b}^*)),$$

where $P_X = X^\top(XX^\top)^{-1}X$. Since $X^\top = P_X X^\top$, we get

$$\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k^{\mathcal{G}}(\mathbf{b}^* + P_X(\mathbf{b}^t - \mathbf{b}^*) + (I-P_X)\boldsymbol{\epsilon}).$$

Let $I^t := \mathrm{supp}(\mathbf{b}^t)\cup\mathrm{supp}(\mathbf{b}^*)$, $\boldsymbol{\lambda}^t := (XX^\top)^{-1}X(\mathbf{b}^t - \mathbf{b}^*)$, and $\mathbf{g} := (I-P_X)\boldsymbol{\epsilon}$. Since $k \geq k^*$, using the properties of the hard thresholding step gives us

$$\left\|\mathbf{b}^{t+1}_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}})\right\|_2 \leq \left\|\mathbf{b}^*_{I^{t+1}} - (\mathbf{b}^*_{I^{t+1}} + X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}})\right\|_2 = \left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}}\right\|_2.$$

This, upon applying the triangle inequality, gives us

$$\left\|\mathbf{b}^{t+1}_{I^{t+1}} - \mathbf{b}^*_{I^{t+1}}\right\|_2 \leq 2\left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t + \mathbf{g}_{I^{t+1}}\right\|_2.$$

Now, using the SGSC and SGSS properties of X (since G-supp$(I^{t+1}) \leq k+k^*$), we can show that $\left\|X^\top_{I^{t+1}}\boldsymbol{\lambda}^t\right\|_2 = \left\|X^\top_{I^{t+1}}(XX^\top)^{-1}X^\top_{I^t}(\mathbf{b}^t - \mathbf{b}^*)\right\|_2 \leq \frac{\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}}\|\mathbf{b}^t - \mathbf{b}^*\|_2$.

Since $\boldsymbol{\epsilon}$ is a Gaussian vector, using tail bounds for Chi-squared random variables (for example, see [5, Lemma 20]), for any set S of size $(k+k^*)d$, we have with probability at least $1-\delta$,$\|\boldsymbol{\epsilon}_S\|_2^2 \leq \sigma^2(k+k^*)d + 2e\sigma^2\sqrt{6(k+k^*)d\log\frac{1}{\delta}}$. Taking a union bound over all sets of group size $(k+k^*)$ and $\binom{n/d}{k} \leq \left(\frac{en}{kd}\right)^k$ gives us, with probability at least $1-\delta$, for all sets S of group size at most $(k+k^*)$,

$$\|\boldsymbol{\epsilon}_S\|_2 \leq \sigma\sqrt{(k+k^*)}\sqrt{d + 2e\sqrt{6d\log\frac{en}{\delta(k+k^*)d}}}$$

From Lemma 24, with probability at least $1-\delta$, we have $\|X\boldsymbol{\epsilon}\|_2 \le 2\sigma\sqrt{n\sqrt{\log n}c'd\log\frac{2d}{\delta}}$. This allows us to bound $\|\mathbf{g}_{I^{t+1}}\|_2$

$$\begin{aligned}
\|\mathbf{g}_{I^{t+1}}\|_2 &= \left\|\boldsymbol{\epsilon}_{I^{t+1}} - X_{I^{t+1}}^\top (XX^\top)^{-1}X\boldsymbol{\epsilon}\right\|_2 \\
&\le \sigma\sqrt{(k+k^*)}\sqrt{d+2e\sqrt{6d\log\frac{en}{\delta(k+k^*)d}}} + 2\sigma\frac{\sqrt{\Lambda_{k+k^*}}}{\lambda_{\frac{n}{d}}}\sqrt{n\sqrt{\log n}c'd\log\frac{2d}{\delta}} \\
&\le \underbrace{\sigma\sqrt{(k+k^*)}\sqrt{d+2e\sqrt{6d\log\frac{en}{\delta(k+k^*)d}}}}_{e_0}\left(1+2f(\mathbf{w}^*,\sigma)\sqrt{\frac{c'd\,(\log n)^{3/2}}{n}\log\frac{2d}{\delta}}\right) \\
&= 1.0003e_0,
\end{aligned}$$

where the second last step is due to Remark 23 for sufficiently large enough n so that $\sqrt{n} \ll n$. Note that e_0 does note depend on the iterates and is thus, a constant. This gives us

$$\left\|\mathbf{b}^{t+1}-\mathbf{b}^*\right\|_2 \le \frac{2\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}}\left\|\mathbf{b}^t-\mathbf{b}^*\right\|_2 + 2.0006e_0.$$

For data matrices sampled from AO-AR(d) ensembles, whose SGSC and SGSS properties are established in Theorem 22, assuming $n \ge d\log d$, we have $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)d\log\frac{n}{\delta(k+k^*)d}}\right)$. Thus, if $\frac{\Lambda_{k+k^*}}{\lambda_{\frac{n}{d}}} < \frac{1}{4}$ (which is guaranteed by Remark 23 with probability at least $1-\delta$ for $k^* \le k \le C\frac{\mathfrak{m}_{\mathbf{w}^*}}{\mathcal{M}_{\mathbf{w}^*}+\mathcal{M}_W}\frac{n}{d\log n}$), then in $T_0 = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{e_0}\right) = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{\sqrt{n}}\right)$ steps, CRTSE ensures that $\left\|\mathbf{b}^{T_0}-\mathbf{b}^*\right\|_2 \le 4.0015e_0$.

□

B.8 Fine Convergence Analysis

Lemma 11. *Suppose $k^* \le k \le n/(C'd\log n)$ for some large enough constant C'. Then with probability at least $1-\delta$,* CRR *ensures at every time instant $t > T_0$*

$$\frac{C}{\lambda_n}(1+\frac{\Lambda_n}{\lambda_n})\left\|X_{FA^{t+1}}(X_{FA^{t+1}}^\top\boldsymbol{\lambda}^t+\mathbf{g}_{FA^{t+1}})\right\|_2 \le 0.5\left\|\boldsymbol{\lambda}^t\right\|_2 + \mathcal{O}\left(\sigma\sqrt{\frac{d\log n}{n}\log\frac{1}{\delta}}\right)$$

Proof. As before, we change the problem so that instead of thresholding the top k elements of the vector $X^\top\boldsymbol{\lambda}^t+\mathbf{g}$ by magnitude, we threshold all elements which exceed a certain value τ in magnitude. Again as before, we show that with high probability, for sufficiently small k, the k^{th} largest element of the vector will have a large magnitude.

Proving the second part of the result is relatively simple in the time series setting because of the error tolerance bound $k^* \le n/(d\log n)$ that we assume in this setting. For sake of simplicity, as well as without loss of generality, assume as before that $\sigma = 1$. Then using the tail bounds for martingales with sub-Gaussian entries from [19], we can yet again show that with probability at least $1-\exp(\Omega(n))$, at least a $1/50$ fraction of points in the vector $\mathbf{g}$ will exceed the value 1.75 in magnitude.

Now, using the subset smoothness of the data matrix X from Theorem 22 on subsets of size 1 tells us that $\max_i\|X_i\|_2 \le \Lambda_1 \le \mathcal{O}(d\log n)$, where X_i is the i^{th} column of the data matrix X. Note that this also includes the influence of the error vector $\mathbf{e}^*$. Thus, if we assume $k^* \le k < \mathcal{O}\left(\frac{1}{d\log n}\right)$ then $\|\boldsymbol{\lambda}^{T_0}\|_2 \le \frac{1}{4\max_i\|X_i\|_2}$ which gives us $\left\|X^\top\boldsymbol{\lambda}^t\right\|_\infty \le \frac{1}{4}$. This assures us that for any $k < n/50$, the k largest elements by magnitude in the vector $X^\top\boldsymbol{\lambda}^t+\mathbf{g}$ will be larger than 1.5.

Having assured ourselves of this, we move on to the analysis assuming that thresholding is done by value and not by cardinality. Let $\mathbf{z} = [z_1, z_2, \ldots, z_n]$ where $z_i = (X_i^\top\boldsymbol{\lambda}+g_i)\cdot\mathbb{I}\left\{\left|X_i^\top\boldsymbol{\lambda}+g_i\right| > \tau\right\}$. We have

$$X\mathbf{z} = \sum_{j=1}^n X_j z_j = \sum_{i=1}^n X_j(X_j^\top\boldsymbol{\lambda}+g_j)\cdot\mathbb{I}\left\{\left|X_j^\top\boldsymbol{\lambda}+g_j\right| > \tau\right\},$$

where the previous result ensures that we can set $\tau \geq 1.5$, as well as safely assume that $\left|X_j^\top \boldsymbol{\lambda}\right| \leq 0.25$. In the following, we analyze the i^{th} coordinate of the vector i.e.

$$(X\mathbf{z})_i = \sum_{i=1}^{n} X_j^i (X_j^\top \boldsymbol{\lambda} + g_j) \cdot \mathbb{I}\left\{\left|X_j^\top \boldsymbol{\lambda} + g_j\right| > \tau\right\} =: \sum_{i=1}^{n} \zeta_i.$$

We notice that $g_i | X_i \sim \mathcal{N}(0, \sigma^2)$ which allows us to construct the following martingale difference sequence

$$\sum_{i=1}^{n} \zeta_i - \mathbb{E}\left[\zeta_i | g_1, g_2, \ldots, g_{i-1}\right]$$

We also note that the elements of the above sequence are conditionally sub-Gaussian with the sub-Gaussian norm at most $\mathcal{O}(\log n)$. Then using the Azuma style inequality for martingales with sub-Gaussian tails from [19] gives us, with high probability

$$\sum_{i=1}^{n} \zeta_i - \mathbb{E}\left[\zeta_i | g_1, g_2, \ldots, g_{i-1}\right] \leq \sqrt{\frac{112 \log n}{n} \log \frac{1}{\delta}}$$

Note that $\mathbb{E}\left[\zeta_i | g_1, g_2, \ldots, g_{i-1}\right] = X_j^i \cdot \mathbb{E}\left[(X_j^\top \boldsymbol{\lambda} + g_j) \cdot \mathbb{I}\left\{\left|X_j^\top \boldsymbol{\lambda} + g_j\right| > \tau\right\} | g_1, g_2, \ldots, g_{i-1}\right]$. Also note that $(X_j^\top \boldsymbol{\lambda} + g_j)$ is conditionally distributed as $\mathcal{N}(X_j^\top \boldsymbol{\lambda}, 1)$ as we have assumed $\sigma = 1$ for simplicity. For a Gaussian variable $Y \sim \mathcal{N}(\mu, 1)$, we have

$$\mathbb{E}\left[Y \cdot \mathbb{I}\left\{|Y| > \tau\right\}\right] = \mu - \mathbb{E}\left[Y \cdot \mathbb{I}\left\{|Y| \leq \tau\right\}\right] = \frac{\phi(\tau - \mu) - \phi(-\tau - \mu)}{\Phi(-\tau - \mu) - \Phi(\tau - \mu)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively, the density and cumulative distribution functions of the standard normal variable. Now, applying the mean value theorem gives us

$$|\phi(\tau - \mu) - \phi(-\tau - \mu)| = |\phi(\tau + \mu) - \phi(\tau - \mu)| = 2\left|\eta \phi(\eta) \mu\right|,$$

for some $\eta \in [\tau - \mu, \tau + \mu]$. For the ensured values of $\tau = 1.25$ and $|\mu| \leq 0.25$, we have $|\phi(\tau - \mu) - \phi(-\tau - \mu)| < 0.25$. For the same values we have $\Phi(-\tau - \mu) - \Phi(\tau - \mu) \geq 0.68$. Putting these together, we get

$$|(X\mathbf{z})_i| \leq C\tau \cdot \left|\sum_{j=1}^{n} X_j^i X_j^\top \boldsymbol{\lambda}\right| + D$$

where $|C_\tau| \leq 0.4$ and $D \leq \sqrt{\frac{112 \log n}{n} \log \frac{1}{\delta}}$. We note that this value of C_τ can be made arbitrarily small by simply requiring that $k^* \leq k < \frac{n}{C' \cdot d \log n}$ for a large enough constant $C' > 0$. In particular, we set k, k^* such that $C_\tau \leq 0.9 \frac{\Lambda_n}{\lambda_n}$. This gives us

$$\frac{1}{\lambda_n} \|X\mathbf{z}\|_2 \leq \frac{C_\tau}{\lambda_n} \left\|XX^\top \boldsymbol{\lambda}\right\|_2 + \frac{d}{\lambda_n} D \leq 0.5 \|\boldsymbol{\lambda}\|_2 + \mathcal{O}\left(\sqrt{\frac{d \log n}{n} \log \frac{1}{\delta}}\right),$$

which concludes the proof. □