FINITE SAMPLE PERFORMANCE OF LINEAR LEAST SQUARES ESTIMATORS UNDER SUB-GAUSSIAN MARTINGALE DIFFERENCE NOISE

Michael Krikheli [*]

Faculty of Engineering,
Bar-Ilan University 52900,
Ramat-Gan Israel,
michael.krih@gmail.com

Amir Leshem

Faculty of Engineering,
Bar-Ilan University 52900,
Ramat-Gan Israel

ABSTRACT

Linear Least Squares is a very well known technique for parameter estimation, which is used even when sub-optimal, because of its very low computational requirements and the fact that exact knowledge of the noise statistics is not required. Surprisingly, bounding the probability of large errors with finitely many samples has been left open, especially when dealing with correlated noise with unknown covariance. In this paper we analyze the finite sample performance of the linear least squares estimator under sub-Gaussian martingale difference noise. In order to analyze this important question we used concentration of measure bounds. When applying these bounds we obtained tight bounds on the tail of the estimator's distribution. We show the fast exponential convergence of the number of samples required to ensure a given accuracy with high probability. We provide probability tail bounds on the estimation error's norm. Our analysis method is simple and uses simple L_∞ type bounds on the estimation error. The tightness of the bounds is tested through simulation. The proposed bounds make it possible to predict the number of samples required for least squares estimation even when least squares is sub-optimal and used for computational simplicity. The finite sample analysis of least squares models with this general noise model is novel.

Index Terms— Estimation; linear least squares; non-Gaussian; concentration bounds; finite sample; large deviations; confidence bounds; martingale difference sequence

1. INTRODUCTION

1.1. Related Work

Linear least squares estimation has numerous applications in many fields. For instance, it was used in soft-decision image interpolation applications in [1] and [2]. Another field that uses linear least squares is source localization using signal strength, as in [3]. In that paper, weighted linear least squares was used to find the distance of the received signals given the strength of the signals received in the sensors and the sensors' locations. Weighted least squares estimators were also used in the field of diffusion MRI parameters estimation [4]. It was shown that the weighted linear least squares approach has significant advantages because of its simplicity and good results. A standard analysis of estimation problems calculates the Cramer-Rao bound (CRB) and uses the asymptotic normality of the estimator.

This type of analysis is asymptotic by nature. For some applications, see for instance [5] where direction of arrival problems were analyzed in terms of the CRB. In [6] the ML estimator and MUSIC algorithms were studied and the CRB was calculated. However, as is well known, the Central Limit Theorem, and the Gaussian approximation are not valid in the case of rare large errors. In many applications the performance is not impacted by small errors, but large errors can lead to catastrophic results. One such example is in wireless communication channel estimation, where the accuracy of the channel estimation should suffice for the given modulation. However rare events where the estimation is significantly far away can lead to total failure. Furthermore, in such applications, training is short and we cannot rely on asymptotic large deviation results. Hence we need tight upper bounds on the L_∞ norm of the error.

The noise model differs accross applications of least squares and other optimization methods. Rather than the Gaussian model a Gaussian mixture is used in many applications. For instance, in [7] a Gaussian mixture model of a time-varying autoregressive process was assumed and analyzed. The Gaussian mixture model was used to model noise in underwater communication systems in [8]. Wiener filters in Guassian mixture signal estimation were analyzed in [9]. In [10] a likelihood based algorithm for Gaussian mixture noise was devised and analyzed in the terms of the CRLB. In [11] a robust detection technique using Maximum-Likelihood estimation was proposed for an impulsive noise modeled as a Gaussian mixture. In this work we consider sub-Gaussian noise, which is a general non-Gaussian noise framework. The Gaussian mixture model, for instance, is sub-Gaussian and our results are valid for this model. In the case of Gaussian noise, least squares coincides with the maximum likelihood estimator. Still, in many cases of interest least squares estimation is used in non-Gaussian noise as well for computational simplicity. Specifically the sub-Gaussian noise model is of special interest in many applications.

In many cases the noise model used is not i.i.d but the noise is correlated. An important case is that of martingale difference noise. This noise model is quite general and is used in various fields. For example the first order ARCH models introduced in [12] are popular in economic theory. Moreover, [13] analyzed similar least squares models with applications in control theory. The asymptotic properties of these models have been analyzed in various papers, for example [14–16]. The results show the strong consistency of the least squares estimator under martingale difference noise and for autoregressive models. The least squares efficiency in an autoregressive noise model was studied in [17]. However, finite sample results were not given.

The least squares problem is well studied. The strong consistency of

the linear least squares was proved in [18]. Asymptotic bounds for fixed size confidence bounds were stated for example in [19]. In the past few years, the finite sample behavior of least squares problems has been studied in [20–23]. Some of these results also analyze regularized least squares models. These results only studied the i.i.d noise case. In this work we extend these results to the sub-Gaussian MDS noise case which is much more general. Beyond the theoretical results we also provide simulated examples of the bounds for the problem of channel estimation with a random mixing matrix.

1.2. Contribution

In this paper we provide a finite sample analysis of linear least squares problems under sub-Gaussian martingale difference sequence (MDS) noise. We provide L_∞ error bounds that can be used to compute the confidence interval in a non-parametric way (i.e., without knowing the exact distribution) of the estimation error. The main theorem of this paper allows us to compute the performance of linear least squares under very general conditions. Since the linear least squares solution is computationally simple it is used in practice even when it is sub-optimal. The analysis of this paper allows the designer to understand the loss due to the computational complexity reduction without the need for massive simulations. We extend the results of [24] in two significant ways. The first is allowing the mixing matrix to be a general bounded elements matrix. More importantly, we extend the analysis to the case of sub-Gaussian MDS noise. The sub Gaussian martingale noise covers many examples of correlated noise, and specifically the case of an interfering zero mean signal which passes through a finite impulse response channel. Hence we are able to predict large error behavior. This provides finite sample analysis under a very general noise framework. While the bounds are not tight, they are still useful and pave the way to further analyses which may tighten these bounds even further. The fact that we only need knowledge of a sub-Gaussianity parameter of the noise allows us to use these bounds when the noise distribution is unknown.

2. PROBLEM FORMULATION

Consider a linear model with additive noise

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{\theta}_0 + \boldsymbol{v} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ is our output, $\boldsymbol{A} \in \mathbb{R}^{N \times p}$ is a known matrix with bounded random elements, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the estimated parameter and $\boldsymbol{v} \in \mathbb{R}^{N \times 1}$ is a noise vector with independent and sub-Gaussian elements[1]. N indicates the number of samples used in the model.
Many real world noise models are sub-Gaussian including Gaussians, finite Gaussian mixtures, all the bounded variables, and any combination of the above. Many real world applications are subject to such noise.
The least squares estimator with N samples is given by

$$\hat{\boldsymbol{\theta}}_0^N = \left(\boldsymbol{A}^T \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{x} = \left(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{a}_n \boldsymbol{a}_n^T\right)^{-1} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{a}_n^T x_n \tag{2}$$

where $\boldsymbol{a}_n^T$, $n = 1 \ldots N$ are the rows of $\boldsymbol{A}$ and x_n, $n = 1 \ldots N$ are the data samples. When $E(\boldsymbol{v}) = \boldsymbol{0}$, $E\left(\hat{\boldsymbol{\theta}}_0^N\right) = \boldsymbol{\theta}_0$ and

---

[1]For simplicity we only consider the real case. The complex case is similar with minor modifications.

the estimator is unbiased.
We want to study the tail distribution of $\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty$ or more specifically to obtain bounds of the form

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) < \varepsilon \tag{3}$$

as a function of N. Furthermore, given r, ε we want to calculate the number of samples needed $N(r, \varepsilon)$ to achieve the above inequality. We analyze the case where $\boldsymbol{A}$ is random with bounded elements. Throughout this paper we use the following mathematical notations:

Definition 2.1.

1. Let $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ be a square matrix; we define the operators $\lambda_{max}(\boldsymbol{B})$ and $\lambda_{min}(\boldsymbol{B})$ to give the maximal and minimal eigenvalues of $\boldsymbol{B}$ respectively.

2. Let $\boldsymbol{C}$ be a matrix. The spectral norm for matrices is given by $\|\boldsymbol{C}\| \doteq \sqrt{\lambda_{max}(\boldsymbol{C}^T \boldsymbol{C})}$.

3. A random variable v with $E(v) = 0$ is called sub-Gaussian if its moment generating function exists and $E(\exp(sv)) \leq \exp\left(\frac{s^2 R^2}{2}\right)$ [25]. The minimal R that satisfies this inequality is called the sub-Gaussian parameter of the random variable v and we say that v is sub-Gaussian with parameter R.

Remark 2.2. Assume that $x \sim N\left(0, \sigma^2\right)$, then the moment generating function of x is $M(s) = E(\exp(sx)) = \exp\left(\frac{s^2 \sigma^2}{2}\right)$. Therefore, by definition 3 x is also sub-Gaussian with parameter σ.

3. MAIN RESULT

In this section we formulate the main result of this paper, discuss it and provide a proof outline.
We make the following assumptions regarding the problem. These assumptions are mild and cover a very large set of linear least squares problems.

A1: $E(v_n) = 0 \quad \forall 1 \leq n \leq N$.

A2: $P\left(\text{rank}\left(\boldsymbol{A}^T \boldsymbol{A}\right) = p\right) = 1$

A3: $P(|a_{ni}| \leq \alpha) = 1 \quad \forall n = 1 \ldots N \quad \forall i = 1 \ldots p$

A4: For all $N > 0$ there exists $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ such that $\boldsymbol{M} = \frac{1}{N}E\left(\boldsymbol{A}^T \boldsymbol{A}\right)$. We denote $\sigma_{max} \doteq \lambda_{max}(\boldsymbol{M})$ and $\sigma_{min} = \lambda_{min}(\boldsymbol{M})$.

A5: $E(v_n | F_{n-1}) = 0$. Where F_{n-1} is a filtration, v_n are independent of $\boldsymbol{A}$.

A6: The martingale difference sequence is δ sub-Gaussian; i.e. $E\left(e^{sv_n} | F_{n-1}\right) \leq e^{\frac{s^2 \delta^2}{2}}$.

Assumptions A1-A2 are standard in least squares theory. Assumption A1 assumes that our design is correct. Assumption A2 ensures that the least squares estimator exists. Assumptions A3 and A4 are mild and achievable by normalizing each row of the mixing matrix with the proper scaling of the sub-Gaussian parameter. Assumption A5 means that the noise sequence is a martingale difference sequence and assumption A6 assumes that the noise sequence is sub-Gaussian. Note that the set of assumptions is valid for any type of martingale difference zero mean sub-Gaussian noise model, which is a very wide family of distributions.

The main theorem provides bounds on the convergence rate of the finite sample least squares estimator to the real parameter. The theorem provides the number of samples needed so that the distance between the estimator and the real parameter will be at most r with probability $1 - \varepsilon$.

Theorem 3.1. (Main Theorem)
Let $\boldsymbol{x}$ be defined as in (1) and assume assumptions A1-A6. Let $\varepsilon > 0$ and $r > 0$ be given and $\hat{\boldsymbol{\theta}}_0^N$ and $\boldsymbol{\theta}_0$ be defined as previously, then $\forall N > N(r, \varepsilon)$

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right) < \varepsilon \qquad (4)$$

where

$$N(r, \varepsilon) = \max\left\{N_1(r, \varepsilon), N_{rand}(\varepsilon)\right\}, \qquad (5)$$

$$N_1(r, \varepsilon) = \frac{8\alpha^2\delta^2}{r^2\sigma_{min}^2} \log \frac{2p}{\varepsilon} \qquad (6)$$

and

$$N_{rand}(\varepsilon) = \frac{4}{3} \frac{(6\sigma_{max} + \sigma_{min})(p\alpha^2 + \sigma_{max})}{\sigma_{min}^2} \log \frac{2p}{\varepsilon}. \qquad (7)$$

where $\sigma_{min}, \sigma_{max}$ have been defined in assumption A4 and δ is the sub-Gaussian parameter and has been defined in assumption A6.

3.1. Discussion

The importance of this result is that it gives an easily calculated bound on the number of samples needed for linear least squares problems. It shows a sharp convergence in probability as a function of N, and shows that the number of samples is $O\left(\frac{1}{r^2} \log \frac{1}{\varepsilon}\right)$. Moreover, the result handles the case where the noise is a martingale difference sequence. This is the first finite sample analysis result for least squares under this noise assumption.

The results in this work are given with an L^∞ norm. The L^∞ results can give confidence bounds for every coordinate of the parameter vector $\boldsymbol{\theta}_0$. Results for other norms can be achieved as well using the relationships between norms.

We start by stating two auxiliary lemmas

Lemma 3.2. *Let $\boldsymbol{x}$ be defined as in (1). Assume A1-A6 hold. Furthermore, let $\hat{\boldsymbol{\theta}}_0^N$ be defined in (2) and let $r > 0$ be given, then*

$$\begin{aligned} &P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right) \\ \leq \quad &P\left(\left|\frac{1}{N}\sum_{n=1}^N a_{ni}v_n\right| > \frac{r}{\lambda_{max}\left(\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\right)}\right). \end{aligned} \qquad (8)$$

Proof. This lemma can be proven by a straightforward computation and the proof is left to the reader. $\qquad \square$

Lemma 3.3. *Under assumptions A2-A4 and for all $N \geq N_{rand}(\varepsilon')$*

$$P\left(\lambda_{max}\left(\frac{1}{N}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\right)^{-1} \geq \frac{2}{\sigma_{min}}\right) \leq \varepsilon', \qquad (9)$$

where

$$N_{rand}(\varepsilon') = \frac{4}{3} \frac{(6\sigma_{max} + \sigma_{min})(p\alpha^2 + \sigma_{max})}{\sigma_{min}^2} \log\left(\frac{p}{\varepsilon'}\right). \qquad (10)$$

Proof. The proof is a generalization of the proof of theorem 4.1 in [24]. It is omitted due to space limitations. $\qquad \square$

3.2. Proof Outline

Proof. We wish to study the term

$$P\left(\left\|\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right\|_\infty > r\right). \qquad (11)$$

In order to do so we start by bounding each of the terms in the vector separately and use a union bound approach to achieve the L_∞ bound. We start by analyzing the term

$$P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right). \qquad (12)$$

Using lemma 3.2 we achieve

$$\begin{aligned} &P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right) \\ \leq \quad &P\left(\left|\frac{1}{N}\sum_{n=1}^N a_{ni}v_n\right| > \frac{r}{\lambda_{max}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}}\right). \end{aligned} \qquad (13)$$

We define the set of events

$$\Psi_1 \doteq \left\{\boldsymbol{X} : \lambda_{max}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)^{-1} \geq \frac{2}{\sigma_{min}}\right\}. \qquad (14)$$

We want to study the number of samples required to achieve that $P(\boldsymbol{X} \in \Psi_1) \leq \frac{\varepsilon}{2}$. In order to achieve this, we use lemma 3.3 with parameter $\varepsilon' = \frac{\varepsilon}{2}$ to find that $\forall N > N_{rand}(\varepsilon)$

$$P(\boldsymbol{X} \in \Psi_1) = P\left(\lambda_{max}\left(\frac{1}{N}\boldsymbol{A}^T\boldsymbol{A}\right)^{-1} \geq \frac{2}{\sigma_{min}}\right) \leq \frac{\varepsilon}{2}. \qquad (15)$$

We denote $\boldsymbol{c}_i$ to be the i-th column of $\boldsymbol{A}$.

We now assume that $\boldsymbol{X} \notin \Psi_1$. Under this assumption the following inequality holds

$$P\left(\left|\left(\hat{\boldsymbol{\theta}}_0^N - \boldsymbol{\theta}_0\right)_i\right| > r\right) \leq P\left(\frac{1}{N}\boldsymbol{c}_i^T\boldsymbol{v} > \frac{r\sigma_{min}}{2}\right). \qquad (16)$$

We denote by

$$\Psi_2(i) \doteq \left\{\boldsymbol{X} : \frac{1}{N}\boldsymbol{c}_i^T\boldsymbol{v} > \frac{r\sigma_{min}}{2}\right\}. \qquad (17)$$

We now obtain a bound on the number of samples required to ensure that

$$P(\boldsymbol{X} \in \Psi_2(i)) = P\left(\frac{1}{N}\boldsymbol{c}_i^T\boldsymbol{v} > \frac{r\sigma_{min}}{2}\right) \leq \frac{\varepsilon}{2p}. \qquad (18)$$

We now outline a proof for a concentration result for a sub-Gaussian martingale difference sequence using similar methods to [26]. We begin by bounding the moment generating function. We start by bounding $E\left(\exp\left(s\boldsymbol{c}_i^T\boldsymbol{v}\right)\right)$ and then we use Markov's inequality. Applying assumption A3 and A6 we obtain

$$E\left(\exp\left(s\boldsymbol{c}_i^T\boldsymbol{v}\right)\right) \leq E\left(\exp\left(s\alpha\sum_{n=1}^{N-1}v_n\right)\right)e^{\frac{s^2\alpha^2\delta^2}{2}}. \qquad (19)$$

Iterating this procedure yields

$$E\left(\exp\left(s\boldsymbol{c}_i^T\boldsymbol{v}\right)\right) \leq \exp\left(\frac{Ns^2\alpha^2\delta^2}{2}\right). \qquad (20)$$

Looking now at the original equation we use the Laplace method and Markov's inequality alongside equation (20) to achieve

$$P\left(\boldsymbol{X} \in \Psi_2\left(i\right)\right) \le \exp\left(\frac{N}{2}\left(s^2\alpha^2\delta^2 - sr\sigma_{min}\right)\right) \qquad (21)$$

Optimizing over $s > 0$ we achieve

$$P\left(\boldsymbol{c}_i^T\boldsymbol{v} > \frac{Nr\sigma_{min}}{2}\right) \le \exp\left(-\frac{Nr^2\sigma_{min}^2}{8\alpha^2\delta^2}\right). \qquad (22)$$

Choosing $N > \frac{8\alpha^2\delta^2}{r^2\sigma_{min}^2}\log\frac{2p}{\varepsilon}$ ensures that

$$P\left(\boldsymbol{X} \in \Psi_2\left(i\right)\right) = P\left(\frac{1}{N}\boldsymbol{c}_i^T\boldsymbol{v} > \frac{r\sigma_{min}}{2}\right) < \frac{\varepsilon}{2p}. \qquad (23)$$

We achieved a bound for each coordinate separately. The last step of the proof is to use the union bound on these terms to achieve a bound on the L_∞ norm of the vector. We define

$$\Psi_2 \doteq \bigcup_{i=1}^{p}\Psi_2\left(i\right). \qquad (24)$$

Using the union bound we obtain $\forall N > N\left(r, \varepsilon\right)$

$$P\left(\boldsymbol{X} \in \Psi_2\right) \le \sum_{i=1}^{p}P\left(\boldsymbol{X} \in \Psi_2\left(i\right)\right) \le \frac{\varepsilon}{2}. \qquad (25)$$

Using the union bound again we obtain $P\left(\boldsymbol{X} \in \Psi_1 \cup \Psi_2\right) \le \varepsilon$. This completes the proof. □

4. SIMULATION RESULTS

Assume that we have a linear system with unknown parameters with noise that is filtered using a finite impulse response system and a sub-Gaussian signal. We can write

$$x_n = \boldsymbol{a}_n^T\boldsymbol{\theta} + \sum_{i=0}^{k}j\left(i\right)H\left(n-i\right) + w\left(n\right). \qquad (26)$$

where $j\left(i\right)$ is an i.i.d zero mean bounded signal for example a BPSK signal. We denote η as the bound for $j\left(i\right)$, i.e. $P\left(j\left(i\right) \le \eta\right) = 1$. We also assume that $H\left(n\right)$ is an unknown system. We now prove that the noise sequence $v_n = \sum_{i=0}^{k}j\left(i\right)H\left(n-i\right) + w\left(n\right)$ is a zero mean martingale difference, that it is sub-Gaussian and thus admits assumptions A5 and A6. If so, we can use theorem 3.1 to calculate the number of samples required to achieve a certain finite sample performance for this interesting model.

$$E\left(v_n|F_{n-1}\right) = E\left(\sum_{i=0}^{k}j\left(i\right)H\left(n-i\right) + w\left(n\right)|F_{n-1}\right)$$

$$= E\left(\sum_{i=0}^{k}j\left(i\right)H\left(n-i\right)|F_{n-1}\right) + E\left(w_n|F_{n-1}\right)$$

$$= E\left(\sum_{i=0}^{k}j\left(i\right)H\left(n-i\right)|F_{n-1}\right) = 0. \qquad (27)$$

The second equality follows from the independence of the random variables $w\left(n\right)$, $j\left(n\right)$ and $H\left(n\right)$. The next equality follows from
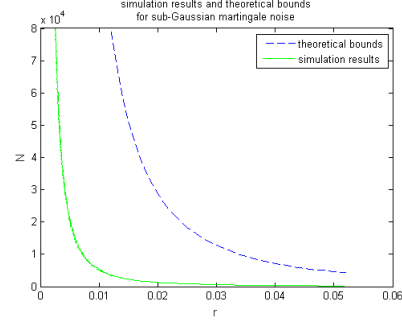

Fig. 1. Simulation results and martingale difference theorem bounds for a sub-Gaussian martingale difference sequence noise with $\varepsilon = 0.2, \sigma_{min} = \sigma_{max} = 10, \delta = 4$ and $p = 2$. The graph is for N as a function of r.

the fact that $w\left(n\right)$ is zero mean. The last equality follows from the fact that $E\left(j\left(n\right)\right) = 0$. We now prove that the v_n is sub-Gaussian. We use the assumption that $j\left(n\right) \le \eta$ and that $H\left(n\right)$ and $w\left(n\right)$ are sub-Gaussian with parameter R_1 and R_2 respectively. Using these facts with the property that $j\left(k\right)H\left(n\right)$ is sub-Gaussian as they are independent and $j\left(i\right) \le \eta$ and the fact that linear combinations of sub-Gaussian random variables is sub-Gaussian [25] we can conclude that v_n is sub-Gaussian and admits assumption A6. We have now proven that this example admits all the assumptions of theorem 3.1 and therefore we can use the theorem to bound the number of samples needed to achieve a predefined performance. Fig 1. shows the performance of the bound in this interesting case. We see that while the bound is not tight, the overall performance is similar. This example demonstrates the strengths of the results in this paper. Many signal processing applications such as this example can be analyzed using our results.

5. CONCLUDING REMARKS

In this paper we examined the finite sample performance of the L^∞ error of the linear least squares estimator. We showed very fast convergence of the number of samples required as a function of the probability of the L^∞ error. We showed that the number of samples required to achieve a maximal deviation r with probability $1 - \varepsilon$ is $N \sim O\left(\frac{1}{r^2}\log\frac{1}{\varepsilon}\right)$. The main theorem deals with least squares in very general noise models; therefore the bounds may be important in many interesting applications. We used simulations to demonstrate the results. Our simulation results suggest that the bounds given in this paper have similar properties as the simulation results. We showed that the interesting example of a finite impulse response filtered interference with sub-Gaussian noise model can be modeled as a sub-Gaussian martingale difference model in our setup and our theorem can give bounds on the number of samples required to achieve the required performance in this important case. This result has significant implications for the analysis of least squares problems in communications and signal processing. We also believe that the Sub-Gaussian parameter can be replaced with bounds on a few moments of the distribution and can relax the bounds. This is left for further study.

6. REFERENCES

[1] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 887–896, June 2008.

[2] K. W. Hung and W. C. Siu, "Robust soft-decision interpolation using weighted least squares," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1061–1069, March 2012.

[3] H. C. So and L. Lin, "Linear least squares approach for accurate received signal strength based source localization," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 4035–4040, Aug 2011.

[4] J. Veraart, J. Sijbers, S. Sunaert, A. Leemans, and B. Jeurissen, "Weighted linear least squares estimation of diffusion MRI parameters: strengths, limitations, and pitfalls," *Neuroimage*, vol. 81, pp. 335–346, 2013.

[5] A. Leshem and A. J. van der Veen, "Direction-of-arrival estimation for constant modulus signals," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 3125–3129, Nov 1999.

[6] P. Stoica and A. Nehorai, "MUSIC, maximum likelihood and Cramer-Rao bound," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr 1988, pp. 2296–2299 vol.4.

[7] P. M. Djurić, J. H. Kotecha, F. Esteve, and E. Perret, "Sequential parameter estimation of time-varying non-Gaussian autoregressive processes," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 865–875, 2002.

[8] S. Banerjee and M. Agrawal, "On the performance of underwater communication system in noise with Gaussian mixture statistics," in *2014 Twentieth National Conference on Communications (NCC)*, Feb 2014, pp. 1–6.

[9] J. Tan, D. Baron, and L. Dai, "Wiener filters in Gaussian mixture signal estimation with ℓ_∞ -norm error," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6626–6635, Oct 2014.

[10] V. Bhatia and B. Mulgrew, "Non-parametric likelihood based channel estimator for Gaussian mixture noise," *Signal Processing*, vol. 87, no. 11, pp. 2569–2586, 2007.

[11] X. Wang and H. V. Poor, "Robust multiuser detection in non-Gaussian channels," *IEEE Transactions on Signal Processing*, vol. 47, no. 2, pp. 289–305, Feb 1999.

[12] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

[13] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, pp. 154–166, 1982.

[14] T. Lai and C. Wei, "Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters," *Journal of multivariate analysis*, vol. 13, no. 1, pp. 1–23, 1983.

[15] P. I. Nelson, "A note on strong consistency of least squares estimators in regression models with martingale difference errors," *The Annals of Statistics*, pp. 1057–1064, 1980.

[16] N. Christopeit and K. Helmes, "Strong consistency of least squares estimators in linear regression models," *The Annals of Statistics*, pp. 778–788, 1980.

[17] W. Krämer, "Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors," *Journal of the American Statistical Association*, vol. 75, no. 372, pp. 1005–1009, 1980.

[18] T. L. Lai, H. Robbins, and C. Z. Wei, "Strong consistency of least squares estimates in multiple regression," in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 510–512.

[19] L. J. Gleser, "On the asymptotic theory of fixed-size sequential confidence bounds for linear regression parameters," *The Annals of Mathematical Statistics*, pp. 463–467, 1965.

[20] R. I. Oliveira, "The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties," *arXiv preprint arXiv:1312.2903*, 2013.

[21] D. Hsu, S. M. Kakade, and T. Zhang, "Random design analysis of ridge regression," *Foundations of Computational Mathematics*, vol. 14, no. 3, pp. 569–600, 2014.

[22] J.-Y. Audibert and O. Catoni, "Robust linear regression through PAC-Bayesian truncation," *Preprint, URL http://arxiv.org/abs/1010.0072*, vol. 38, p. 60, 2010.

[23] ——, "Robust linear least squares regression," *The Annals of Statistics*, pp. 2766–2794, 2011.

[24] M. Krikheli and A. Leshem, "Finite sample performance of least squares estimation in sub-Gaussian noise," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–5.

[25] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[26] G. Thoppe and V. S. Borkar, "A concentration bound for stochastic approximation via Alekseev's formula," *arXiv preprint arXiv:1506.08657*, 2015.