2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Convergence of Parameter Estimates for Regularized Mixed Linear Regression Models *

Taiyao Wang and Ioannis Ch. Paschalidis, *Fellow, IEEE*

***Abstract*— We consider *Mixed Linear Regression (MLR)*, where training data have been generated from a mixture of distinct linear models (or clusters) and we seek to identify the corresponding coefficient vectors. We introduce a *Mixed Integer Programming (MIP)* formulation for MLR subject to regularization constraints on the coefficient vectors. We establish that as the number of training samples grows large, the MIP solution converges to the true coefficient vectors in the absence of noise. Subject to slightly stronger assumptions, we also establish that the MIP identifies the clusters from which the training samples were generated. In the special case where training data come from a single cluster, we establish that the corresponding MIP yields a solution that converges to the true coefficient vector even when training data are perturbed by (martingale difference) noise. We provide a counterexample indicating that in the presence of noise, the MIP may fail to produce the true coefficient vectors for more than one clusters. We also provide numerical results testing the MIP solutions in synthetic examples with noise.**

I. Introduction

Mixed Linear Regression (MLR) [1], [2] is also known as mixtures of linear regressions [3] or cluster-wise linear regression [4]. It involves the identification of two or more linear regression models from unlabeled samples generated from an unknown mixture of these models. This can be seen as a joint clustering and regression problem. The problem is related to the identification of hybrid and switched linear systems [5], [6] and has many diverse applications. In this work, we focus on the fundamental problem of establishing strong consistency of parameter estimates, i.e., establishing that the estimated parameters converge to their true values as the number of the training samples grows.

MLR is typically solved by local search such as *Expectation Maximization (EM)* or alternating minimization, where one alternates between clustering and regression. It has recently been shown that EM converges to the true parameters if it starts from a small enough neighborhood around them [1], [7], [8].

Much effort has focused on the case where training samples are not perturbed by noise. [9] proposed a combination of tensor decomposition and alternating minimization, showing that initialization by the tensor method allows alternating minimization to converge to the global optimum at a linear rate with high probability (w.h.p.). [2] proposes a non-convex objective function with a tensor method for initialization so that the initial coefficients are in the neighborhood of the true coefficients w.h.p. [10] proposes a second-order cone program, showing that it recovers all mixture components in the noiseless setting under conditions that include a well-separation assumption and a balanced measurement assumption on the data. [11] considers data generated from a mixture of Gaussians, showing global convergence of the proposed algorithm with nearly optimal sample complexity.

Establishing convergence (w.h.p.) and consistency results for MLR in the presence of noise is much harder. [3] develops a provably consistent estimator for MLR that relies on a low-rank linear regression to recover a symmetric tensor, which can be factorized into the parameters using a tensor power method. [12] provides a convex optimization formulation for MLR with two components and upper bounds on the recovery errors under subgaussian noise assumptions. [13] studies a *MixLasso* approach but convergence results are limited to the objective function and not the solution. [14] develops an algorithm based on ideas from sparse graph codes; the convergence results however are asymptotic under Gaussian noise and for MLR with only two components.

Recently, an increasing number of machine learning and statistics problems have been tackled using MIP methods [15], [16], [17], [18], [19]. [20] proposes an MIP formulation for MLR and a pre-clustering heuristic approach to solve large-scale problems. [4] proposes a Mixed-Integer Quadratic Program (MIQP) formulation for MLR.

At the same time, *regularization* in learning problems has become widespread, following the early success of LASSO [21], [22]. Recent work has obtained regularization as a consequence of solving a robust learning problem (see, [23] and references therein).

In this paper, we introduce a general MIP formulation for MLR subject to norm-based regularization constraints. We establish that optimal solutions of the MIP converge almost surely (rather than w.h.p.) to the true parameters in the noiseless case as the sample size increases. Subject to cluster separability assumptions, we also establish that MIP solutions can identify the proper cluster for each given sample. For the special case of a single cluster, we show that the MIP solution converges to the true parameter vector in the presence of noise satisfying a martingale difference assumption [24], [25], [26]. For multiple clusters in the

* Research partially supported by the NSF under grants DMS-1664644, CNS-1645681, and IIS-1914792, by the ONR under MURI grant N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, by the Boston University Digital Health Initiative, and by the Boston University Center for Information and Systems Engineering.

[1]Division of Systems Engineering, Boston University, Boston, MA 02446, USA. wty@bu.edu

[2]Dept. of Electrical and Computer Engineering, Division of Systems Engineering, and Dept. of Biomedical Engineering, Boston University, 8 St. Mary's St., Boston, MA 02215, USA. yannisp@bu.edu, http://sites.bu.edu/paschalidis/.

presence of noise we provide a counterexample, suggesting that one can not in general recover the true parameters.

Our convergence analysis leverages techniques for proving strong consistency of least-squares estimates for linear regression, but under weaker assumptions. The related literature is substantial. A breakthrough paper [24] established strong consistency of least-squares estimates for stochastic regression models. Asymptotic properties and strong consistency of least-squares parameter estimates have been studied in many areas including system identification and adaptive control [26], econometric theory [27] and time series analysis [28].

The paper is organized as follows. Sec. II, presents the formulation of the problem. The main results are given in Sec. III. Sec. IV presents numerical results and Sec. V draws conclusions.

Notation: All vectors are column vectors and denoted by bold lowercase letters. Bold uppercase letters denote matrices. For economy of space, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. Prime denotes transpose and $\|\mathbf{x}\|_p = (\sum_{i=1}^{\dim(\mathbf{x})} |x_i|^p)^{1/p}$ the ℓ_p norm, where $p \geq 1$. Unless otherwise specified, $\|\cdot\|$ denotes the ℓ_2 norm and $\|\cdot\|_0$ the ℓ_0 counting "norm." $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the minimum and maximum eigenvalue of a (symmetric) matrix. We use $\emptyset$ to denote the empty set, $[n]$ for the set $\{1, \ldots, n\}$, and $|S|$ for the cardinality of the set S. We write $f(n) = O(g(n))$ if there exist positive numbers n_0 and c such that $f(n) \leq cg(n), \forall n \geq n_0$. We write $f(n) = \Omega(g(n))$ if there exist positive numbers n_0 and c such that $f(n) \geq cg(n), \forall n \geq n_0$. We write $f(n) = \Theta(g(n))$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ hold. Finally, E and P denote expectation and probability, respectively.

II. Problem Formulation

Consider the MLR model where data $(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1}$ are generated by

$$y_i = \mathbf{x}_i' \beta_k + \varepsilon_i, \text{ for some } k \in [K],$$

where $\{\beta_k, \forall k \in [K]\}$ are the ground truth coefficient vectors. The problem is to estimate these parameters from data.

Given a training dataset $\{(\mathbf{x}_i, y_i),\ i \in [n]\}$, we formulate the problem as the following MIP:

$$\begin{aligned} \min_{\beta_k, t_i, c_{ki}} \ & \frac{1}{n} \sum_{i \in [n]} t_i^p \qquad (1) \\ \text{s.t. } & t_i - (y_i - \mathbf{x}_i' \beta_k) + M(1 - c_{ki}) \geq 0,\ i \in [n], k \in [K], \\ & t_i + (y_i - \mathbf{x}_i' \beta_k) + M(1 - c_{ki}) \geq 0,\ i \in [n], k \in [K], \\ & \sum_{k \in [K]} c_{ki} = 1,\ i \in [n], \\ & c_{ki} \in \{0, 1\},\ i \in [n], k \in [K], \\ & t_i \geq 0,\ i \in [n], \\ & \|\beta_k\|_q \leq d_{k,q},\ k \in [K], \end{aligned}$$

where M is a large constant (big-M). Notice that when $c_{ki} = 1$, the first two constraints imply $t_i \geq |(y_i - \mathbf{x}_i' \beta_k)|$, whereas $c_{ki} = 0$ implies that no constraint is imposed on t_i since M is a large positive constant. At optimality, (1) minimizes a p-norm loss function for the regression problem and assigns each data point to the cluster achieving minimal loss. The last constraint in (1) imposes a q-norm regularization constraint to each coefficient vector β_k and $d_{k,q}$ are given constants. In statistics and machine learning, regularization is widely used to help prevent models from overfitting the training data [23]. A Bayesian understanding of regularization is that regularized least squares are equivalent to priors on the solution to the least squares problem. For $p, q = 1$, (1) is a linear MIP, while if either p or q (or both) are equal to 2 it is a quadratic MIP. Both can be solved by modern MIP solvers.

Without the regularization constraint, another equivalent formulation of (1) is

$$\min_{\beta_k} \frac{1}{n} \sum_{i \in [n]} \min_{k \in [K]} |y_i - \mathbf{x}_i' \beta_k|^p. \qquad (2)$$

Let $\{\beta_k^n, t_i^n, c_{ki}^n,\ k \in [K],\ i \in [n]\}$ denote an optimal solution to (1); we use a superscript n to explicitly denote dependence on the training set. Define $\Omega_k^n = \{i \in [n] :\ y_i = \mathbf{x}_i' \beta_k + \varepsilon_i\}$ and $\hat{\Omega}_k^n = \{i \in [n] :\ c_{ki}^n = 1\}$ which form the true and the estimated partition of the training set into the K clusters, respectively. We can show the following lemma (proof omitted due to space limitations).

Lemma II.1 $\forall k \in [K]$,

$$\begin{aligned} \hat{\Omega}_k^n \subset \{i \in [n] :\ |y_i - \mathbf{x}_i' \beta_k^n| = \min_{m \in [K]} |y_i - \mathbf{x}_i' \beta_m^n|\}, \\ \hat{\Omega}_k^n \supset \{i \in [n] :\ |y_i - \mathbf{x}_i' \beta_k^n| < \min_{m \neq k \in [K]} |y_i - \mathbf{x}_i' \beta_m^n|\}. \end{aligned}$$

In order to analyze the strong consistency of parameter estimates obtained by the MIP formulation, we introduce the following assumptions.

(A1) The clusters are not degenerate and different, i.e., $|\Omega_k^n| = \Theta(n)$, $\forall k \in [K]$ and $\beta_k \neq \beta_h$, $\forall k \neq h$.

(A2) The noise sequence $\{\varepsilon_i, \mathscr{F}_i\}$ is a martingale difference sequence, where $\{\mathscr{F}_i\}$ is a sequence of increasing σ-fields, and $\sup_i \mathsf{E}[|\varepsilon_i|^2 | \mathscr{F}_{i-1}] < \infty$, a.s.

(A3) The noise sequence $\{\varepsilon_i, \mathscr{F}_i\}$ is a martingale difference sequence, where $\{\mathscr{F}_i\}$ is a sequence of increasing σ-fields, and $\sup_i \mathsf{E}[|\varepsilon_i|^\alpha | \mathscr{F}_{i-1}] < \infty$, a.s. for some $\alpha > 2$.

(A4) $\|\beta_k\|_q \leq d_{k,q},\ \forall k \in [K],\ \forall q \in \{2, 1, 0\}$.

The noise assumptions are mild. For instance, (A2) holds if $\{\varepsilon_i\}$ are *i.i.d.* random variables with zero mean and variance, including but not limited to Gaussian white noise.

The following lemmata are important in proving the strong consistency of least squares estimates in stochastic linear regression models.

Lemma II.2 *([24]) Suppose the noise sequence $\{\varepsilon_i\}$ satisfies Assumption (A2). Let $\mathbf{x}_i$ be $\mathscr{F}_{i-1}$ measurable for every i. Define $N = \inf\{n :\ \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ is nonsingular $\}$. Assume that $N < \infty$ a.s., and for $n \geq N$, define*

$$Q_n = \left(\sum_{i=1}^n \mathbf{x}_i \varepsilon_i \right)' \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i \varepsilon_i \right).$$

Let $\lambda_{max}(n)$ the maximum eigenvalue of $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$. Then $\lambda_{max}(n)$ is non-decreasing in n and

(i) *On $(\lim_{n\to\infty} \log\lambda_{max}(n) < \infty)$, $Q_n = O(1)$ a.s.*

(ii) *On $(\lim_{n\to\infty} \log\lambda_{max}(n) = \infty)$, we have that for $\forall\delta > 0$,*

$$Q_n = O((\log\lambda_{max}(n))^{1+\delta}) \text{ a.s.}$$

(iii) *On $(\lim_{n\to\infty} \log\lambda_{max}(n) = \infty)$, the previous result can be strengthened to*

$$Q_n = O(\log\lambda_{max}(n)) \text{ a.s.},$$

if the assumption (A2) is replaced by (A3).

The following estimates for the weighted sums of martingale difference sequences are based on the Kronecker lemma, the local convergence theorem and the strong law for martingales [25], [29].

Lemma II.3 *([24], [26]) Suppose the noise sequence $\{\varepsilon_i\}$ satisfies the assumption (A2). Let u_i be $\mathscr{F}_{i-1}$ measurable for every i and $s_n = (\sum_{j\in[n]} u_j^2)^{\frac{1}{2}}$. Then,*

(i)

$$\sum_{j\in[n]} u_j\varepsilon_j \text{ converges a.s. on } s_n^2 < \infty.$$

(ii)

$$\sum_{j\in[n]} u_j\varepsilon_j = o(s_n[log(s_n^2)]^{\frac{1}{2}+\delta}) \text{ a.s. } \forall\delta > 0.$$

$$\text{on } \sum_{j\in[n]} u_j^2 = \infty.$$

(iii)

$$\sum_{j\in[n]} u_j\varepsilon_j = O(s_n[log(s_n^2)]^{\frac{1}{2}}) \text{ a.s. on } \sum_{j\in[n]} u_j^2 = \infty,$$

if the assumption (A2) is replaced by (A3).

The estimates given by these results are not as sharp as those given by the law of the iterative logarithm [25] but the assumptions needed here are much more general.

III. Main Results

A. Noiseless case

Suppose $\{\beta_k^n,\ \forall k \in [K]\}$ are optimal solutions to (1) for $p \in \{2,1\}$ and $q \in \{2,1,0\}$.

Theorem 1 *Suppose Assumptions (A1) and (A4) hold in the MLR model, and*

$$\liminf_{|\mathscr{S}|\to\infty} \lambda_{min}\left(\sum_{i\in\mathscr{S}\subset\Omega_k^n} \mathbf{x}_i\mathbf{x}_i'\right) > 0,\ a.s.,\ \forall k \in [K]. \quad (3)$$

Then we have the strong consistency, i.e.,

$$\exists N \text{ s.t. } \beta_k^n = \beta_{\pi(k)},\ a.s., \forall n > N, k \in [K],$$

where π is a permutation of the K clusters.

Furthermore, if the optimal solution of (1) is unique, or there are no ties for assigning samples to clusters, i.e., $\forall i \in [n], \forall k \neq h \in [K],\ |\mathbf{x}_i'(\beta_k - \beta_h)| > 0$, it follows

$$\exists N \text{ s.t. } \hat{\Omega}_k^n = \Omega_{\pi(k)}^n,\ a.s.,\ \forall n > N. \quad (4)$$

Proof: $\forall k \in [K],\ \exists m(k) \in [K]$ such that

$$|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq |\Omega_k^n \cap \hat{\Omega}_h^n|,\ \forall h \in [K].$$

Consequently,

$$|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq |\Omega_k^n|/K = \Theta(n) \to \infty,\ \text{when } n \to \infty.$$

From (3), $\exists N$, such that $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| > N$, and

$$\lambda_{min}\left(\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} \mathbf{x}_i\mathbf{x}_i'\right) > 0,\ \text{a.s.} \quad (5)$$

Since the true $\{\beta_k\}$ are a feasible solution of (1) and there is no noise, the optimal objective value must be 0, i.e., $y_i = \mathbf{x}_i'\beta_k$ for some k and all i. It follows

$$y_i = \mathbf{x}_i'\beta_k = \mathbf{x}_i'\beta_{m(k)}^n, \quad \forall i \in \Omega_k^n \cap \hat{\Omega}_{m(k)}^n.$$

As a result,

$$\mathbf{x}_i'(\beta_k - \beta_{m(k)}^n) = 0,$$

and

$$\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} |\mathbf{x}_i'(\beta_k - \beta_{m(k)}^n)|^2 = 0. \quad (6)$$

From the definition of the smallest eigenvalue, we have

$$\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} |\mathbf{x}_i'(\beta_k - \beta_{m(k)}^n)|^2 \geq \lambda_{min}\left(\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} \mathbf{x}_i\mathbf{x}_i'\right)\|\beta_k - \beta_{m(k)}^n\|^2. \quad (7)$$

Accordingly, when $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| > N$, it follows from equations (5), (6) and (7) that $\beta_k - \beta_{m(k)}^n = 0$.

Next, we show the mapping $m(\cdot)$ is a bijection by contradiction. Assume there exists $m(k) = m(h)$ for $k \neq h$. Then, $\beta_k = \beta_{m(k)}^n = \beta_{m(h)}^n = \beta_h$ when n is large enough, which contradicts the assumption that the clusters are different. Thus, $\pi = m^{-1}$ is a permutation. Finally, we can prove the cluster set equality (4) by contradiction. ■

Remark 1 *(i) Equation (3) on $\mathbf{x}_i$ is not only sufficient but also necessary. $\mathscr{S}$ in (3) is the subset of Ω_k^n. For instance, if $\mathbf{x}_i = (1,1),\ \forall i \in \Omega_k^n$, and $\lambda_{min}(\sum_{i\in\Omega_k^n} \mathbf{x}_i\mathbf{x}_i^{'}) = 0$, which violates Equation (3), and model parameters are not identifiable no matter which method is being used.*

(ii) Thm. 1 holds for either $p = 1$ or $p = 2$. From a computational complexity aspect, a linear MIP ($p = 1, q = 1$) can be solved faster than a quadratic MIP ($p = 2$).

(iii) Even if we have the strong consistency for the model parameters, i.e., $\beta_k^n \to \beta_{\pi(k)},\ a.s., \forall k \in [K]$, we may not

have $\hat{\Omega}_k^n \to \Omega_{\pi(k)}^n$, *a.s. when* $n \to \infty$. *For instance, if* $\mathbf{x}_i = \mathbf{0}$, *for some i, then the sample i may be assigned to any cluster.*

Thm. 1 holds for any $p > 0$ and $q \geq 0$. Assumption (3) is equivalent to requiring that every cluster has at least d linearly independent measurements. The exact convergence can also be achieved by other methods, e.g., the second-order cone program in [10]. However, to the best of our knowledge, our assumptions are the weakest since we do not need a "sufficient-separation" and a balanced measurement assumption on the data as in [10].

B. Noisy case with a single cluster

Linear regularized regression can be seen as a specific case of MLR with a single cluster. In this section, we assume $K = 1$ and consider the presence of noise. We establish strong consistency for the model parameters under general covariate and noise distributions. Consider the regularized regression problem

$$\min_{\beta} \frac{1}{n} \sum_{i\in[n]} |y_i - \mathbf{x}_i'\beta|^2 \quad \text{s.t. } \|\beta\|_q \leq d_q, \quad \forall k \in [K], \tag{8}$$

where $q \in \{2,1,0\}$, corresponding to ridge regression, LASSO and regression with a subset selection constraint, respectively. For $q = 0$, the problem can be formulated as a MIP problem [16]. Let β^n denote an optimal solution of (8).

Let $\lambda_{max}(n) = \lambda_{max}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')$ and $\lambda_{min}(n) = \lambda_{min}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')$.

Theorem 2 *Suppose that Assumptions (A1), (A2) and (A4) hold for (8), and let* $\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$ *measurable for every i. If*

$$\lambda_{min}(n) \to \infty, \ a.s.,$$

then for all $\delta > 0$,

$$\|\beta^n - \beta\| \leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}(n)), \ a.s.$$

Proof: Since β is a feasible solution of (8) (cf. Ass. (A4)), we have

$$\sum_{i\in[n]} |y_i - \mathbf{x}_i'\beta^n|^2 \leq \sum_{i\in[n]} |y_i - \mathbf{x}_i'\beta|^2 = \sum_{i\in[n]} |\varepsilon_i|^2.$$

Substituting $y_i = \mathbf{x}_i'\beta + \varepsilon_i$ in the l.h.s., we obtain

$$\sum_{i\in[n]} |\varepsilon_i - \mathbf{x}_i'(\beta^n - \beta)|^2 \leq \sum_{i\in[n]} |\varepsilon_i|^2,$$

and by expanding the l.h.s. we obtain

$$\sum_{i\in[n]} |\mathbf{x}_i'(\beta^n - \beta)|^2 \leq 2\sum_{i\in[n]} \varepsilon_i\mathbf{x}_i'(\beta^n - \beta). \tag{9}$$

From the definition of the minimum eigenvalue, we have

$$\sum_{i\in[n]} |\mathbf{x}_i'(\beta^n - \beta)|^2 \geq \lambda_{min}(n)\|\beta^n - \beta\|^2. \tag{10}$$

Next, we study the r.h.s. of (9) in order to bound the convergence rate of $\|\beta^n - \beta\|$. From Lemma II.3(ii), we have that for any $j \in [d]$,

$$\sum_{i\in[n]} \varepsilon_i x_{ij} = o(s_{nj}[log(s_{nj}^2)]^{\frac{1}{2}+\delta}) \text{ a.s., where } s_{nj} = \left(\sum_{i\in[n]} x_{ij}^2\right)^{\frac{1}{2}}.$$

For any $j \in [d]$,

$$s_{nj} = \left(\sum_{i\in[n]} x_{ij}^2\right)^{\frac{1}{2}} \leq \left(\sum_{i\in[n]}\sum_{j\in[d]} x_{ij}^2\right)^{\frac{1}{2}} \leq \left(\mathrm{Tr}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)\right)^{\frac{1}{2}} = \Theta\left(\lambda_{max}^{\frac{1}{2}}(n)\right),$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. By combining the above two equations, we have for any $j \in [d]$,

$$\sum_{i\in[n]} \varepsilon_i x_{ij} = o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \text{ a.s.},$$

and thus,

$$\left\|\sum_{i\in[n]} \varepsilon_i\mathbf{x}_i\right\| = o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \text{ a.s.}$$

We bound the r.h.s. of (9) as

$$2\sum_{i\in[n]} \varepsilon_i\mathbf{x}_i'(\beta^n - \beta) \leq 2\left\|\sum_{i\in[n]} \varepsilon_i\mathbf{x}_i\right\| \|\beta^n - \beta\| \leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta})\|\beta^n - \beta\|. \tag{11}$$

Combining (9), (10) and (11), we obtain

$$\lambda_{min}(n)\|\beta^n - \beta\|^2 \leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta})\|\beta^n - \beta\|.$$

■

If the Assumption (A2) is replaced by (A3), we have a stronger version of the previous theorem (proof omitted).

Theorem 3 *Suppose that Assumptions (A1), (A3) and (A4) hold for (8), and let* $\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$ *measurable for every i. If*

$$\lambda_{min}(n) \to \infty, \ a.s.,$$

then we have

$$\|\beta^n - \beta\| \leq O(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}}/\lambda_{min}(n)) \ a.s.$$

As a point of comparison, the classical convergence rate for least square estimates of unregularized linear regression subject to martingale difference noise is given by:

$$\|\beta^n - \beta\| = o([log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}^{\frac{1}{2}}(n)) \text{ a.s. for } \alpha = 2, \quad = O([log\lambda_{max}(n)]^{\frac{1}{2}}/\lambda_{min}^{\frac{1}{2}}(n)) \text{ a.s. for } \alpha > 2. \tag{12}$$

The convergence rate in Theorem 2 and Theorem 3 is slower than the unregularized linear regression case in general. However, it can be seen that for well-conditioned data, i.e., $\lambda_{max}(n) \sim \lambda_{min}(n)$, we have the same convergence rate with

the unregularized case. The following Corollary outlines sufficient conditions to that effect.

Corollary 1 *Suppose* $\{\mathbf{x}_i\}$ *are i.i.d. random vectors with* $\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']$ *being a positive definite matrix. Suppose* $\{\varepsilon_i\}$ *are i.i.d. random variables, independent of the* $\mathbf{x}_i$*, with zero mean and variance* $\sigma^2 > 0$*. Then, we have*

$$\|\beta^n - \beta\|^2 = o(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}+\delta}),\ a.s.\ \forall \delta > 0.$$

Furthermore, if $\mathsf{E}[|\varepsilon_i|^\alpha] < \infty$*, a.s. for some* $\alpha > 2$*, we have*

$$\|\beta^n - \beta\|^2 = O(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}}),\ a.s. \tag{13}$$

Proof: From the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}(\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i') = \mathsf{E}[\mathbf{x}_i\mathbf{x}_i'],\ \text{a.s.}$$

It follows

$$\lambda_{max}(n) = \Theta(n), \qquad \lambda_{min}(n) = \Theta(n).$$

Thus, from Thm. 2 we have

$$\begin{aligned}\|\beta^n - \beta\| \leq & o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}(n)) \\ = & o(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}+\delta}),\ \text{a.s.}\ \forall \delta > 0.\end{aligned}$$

Similarly, using Thm. 3 we can prove (13). ■

We point out that our setting is more general, including ridge regression, LASSO and regression with subset selection. The convergence rate is sharper in terms of the sample size than the rate achieved for LASSO in [22]. In addition, our noise assumptions are more general and weaker than the Gaussian noise used in [22].

C. A counterexample for the noisy case with two clusters

Next, we provide a counterexample indicating that the presence of noise may prevent convergence to the true coefficients when we have more than a single cluster.

Consider the 2-cluster MLR model where the data $(x_i, y_i) \in \mathbb{R}^2$ are generated as follows:

$$\begin{aligned} x_i =& 1, \\ \varepsilon_i \in& \{1,-1\} \text{ with probability 0.5,} \\ \beta_k \in& \{\delta, 0\} \text{ with probability 0.5,} \\ y_i =& x_i\beta_k + \sigma\varepsilon_i \in \{\sigma, -\sigma, \delta+\sigma, \delta-\sigma\}, \end{aligned}$$

where $\{\varepsilon_i\}$ are i.i.d. random variables with zero mean and positive variance. If $\delta < \sigma$, the optimal solution of (1) will be different from the ground truth parameters. In particular, the optimal objective function value of (1), is smaller than the objective function value of the ground truth parameters, and depend on δ rather than σ.

Not only (1), but also other methods may fail to recover the ground truth parameters since they are "unidentifiable," i.e., two sets of parameters can generate the same distribution. This counterexample shows strong consistency may fail to hold under the weakest assumptions used in our earlier analysis.

IV. Numerical results

In this section we provide numerical results on the convergence of parameters estimates in MLR obtained by formulation (1). Computations were performed with GUROBI 8.0 and python 3.6.5 with 16 CPUs on the Boston University Shared Computing Cluster. The results below coincide with the intuition that if the K clusters are well-separated, the convergence of the estimates becomes almost equivalent to having K separate linear regression models.

A. MLR under Gaussian noise

Consider a model where the data $(\mathbf{x}_i, y_i) \in \mathbb{R}^3$ are generated independently by

$$\begin{aligned} x_{i,1} \sim& UNIFORM(0,1), \quad x_{i,2} = 1, \\ \beta_1 =& (-0.93, 0.1), \quad \beta_2 = (0,0), \\ \varepsilon_i \sim& N(0,1), \\ y_i =& \mathbf{x}_i'\beta_k + 0.01\varepsilon_i,\ \text{for some } k \in [2]. \end{aligned}$$

The first element of of $\mathbf{x}_i$ consists of random samples from the uniform distribution over $[0,1)$. The second element of $\mathbf{x}_i$ is constant. ε_i are drawn from the standard normal distribution. One-half of the data samples are from Cluster 1. Fig. 1 plots the evolution of $\|\beta_k^n - \beta_k\|$ as a function of the number of samples n used in solving problem (1), where we used $p = 1$, $M = 10$, did not include a regularization constraint, and set the time limit for each model in GUROBI to 2 hours.

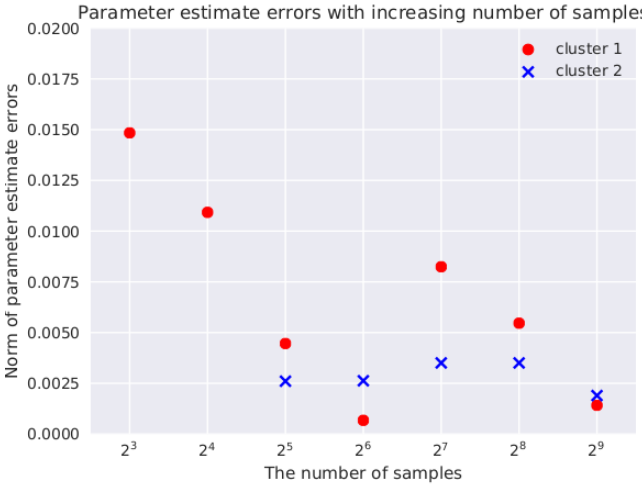

Fig. 1: Gaussian noise case.

B. MLR under uniform noise

Consider now a model where the data samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^3$ are generated independently by

$$\begin{aligned} x_{i,1} \sim& UNIFORM(0,1), \quad x_{i,2} = 1, \\ \beta_1 =& (-1.61, 1.25), \quad \beta_2 = (0,0), \\ \varepsilon_i \sim& UNIFORM(-1,1), \\ y_i =& \mathbf{x}_i'\beta_k + 0.01\varepsilon_i,\ \text{for some } k \in [2]. \end{aligned}$$

The first element of $\mathbf{x}_i$ consists of random samples from the uniform distribution over $[0, 1)$, and the second element is the constant 1. ε_i are drawn from the uniform distribution over $[-1, 1)$. One-half of the samples are generated from Cluster 1. Fig. 2 plots $\|\beta_k^n - \beta_k\|$ as a function of the number of samples used in (1).

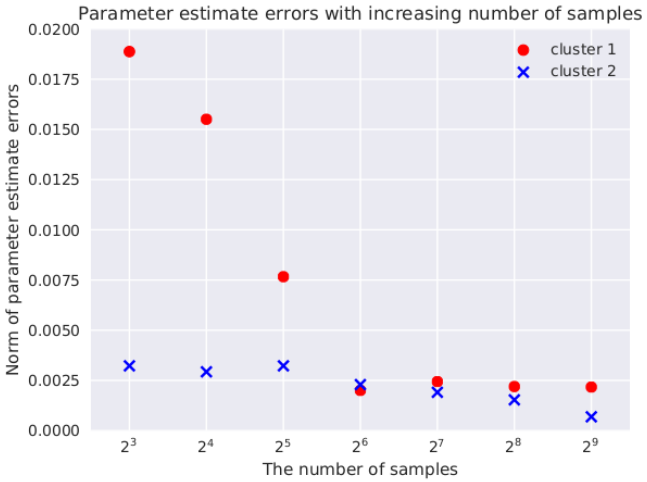

Fig. 2: Uniform noise case.

V. Conclusions

We established the convergence of parameter estimates for regularized mixed linear regression models with multiple components in the noiseless case. Regularized linear regression can be seen as a specific case of MLR with a single cluster. We establish strong consistency and characterize the convergence rate of parameter estimates for such regression models subject to martingale difference noise under very weak assumptions on the data distribution.

To the best of our knowledge, our paper is the first to study strong consistency of parameter estimates for mixed linear regression models under general noise conditions and general feature conditions rather than convergence with high probability. It can be used directly and extended in many areas, including but not limited to system identification and control, econometric theory and time series analysis.

References

[1] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," in *International Conference on Machine Learning*, 2014, pp. 613–621.

[2] K. Zhong, P. Jain, and I. S. Dhillon, "Mixed linear regression with multiple components," in *Advances in neural information processing systems*, 2016, pp. 2190–2198.

[3] A. T. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," in *International Conference on Machine Learning*, 2013, pp. 1040–1048.

[4] Y. W. Park, Y. Jiang, D. Klabjan, and L. Williams, "Algorithms for generalized clusterwise linear regression," *INFORMS Journal on Computing*, vol. 29, no. 2, pp. 301–317, 2017.

[5] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems a tutorial," *European journal of control*, vol. 13, no. 2-3, pp. 242–260, 2007.

[6] R. Vidal, "Recursive identification of switched arx systems," *Automatica*, vol. 44, no. 9, pp. 2274–2287, 2008.

[7] X. Yi and C. Caramanis, "Regularized em algorithms: A unified framework and statistical guarantees," in *Advances in Neural Information Processing Systems*, 2015, pp. 1567–1575.

[8] S. Balakrishnan, M. J. Wainwright, B. Yu *et al.*, "Statistical guarantees for the em algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[9] X. Yi, C. Caramanis, and S. Sanghavi, "Solving a mixture of many random linear equations by tensor decomposition and alternating minimization," *arXiv preprint arXiv:1608.05749*, 2016.

[10] P. Hand and B. Joshi, "A convex program for mixed linear regression with a recovery guarantee for well-separated data," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 563–579, 2018.

[11] Y. Li and Y. Liang, "Learning mixtures of linear regressions with nearly optimal complexity," *arXiv preprint arXiv:1802.07895*, 2018.

[12] Y. Chen, X. Yi, and C. Caramanis, "A convex formulation for mixed regression with two components: Minimax optimal rates," in *Conference on Learning Theory*, 2014, pp. 560–604.

[13] I. E.-H. Yen, W.-C. Lee, K. Zhong, S.-E. Chang, P. K. Ravikumar, and S.-D. Lin, "Mixlasso: Generalized mixed regression via convex atomic-norm regularization," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 891–10 899.

[14] D. Yin, R. Pedarsani, Y. Chen, and K. Ramchandran, "Learning mixtures of sparse linear regressions using sparse graph codes," *IEEE Transactions on Information Theory*, 2018.

[15] D. Bertsimas and A. King, "Or forum—an algorithmic approach to linear regression," *Operations Research*, vol. 64, no. 1, pp. 2–16, 2015.

[16] D. Bertsimas, A. King, R. Mazumder *et al.*, "Best subset selection via a modern optimization lens," *The annals of statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[17] T. Xu, T. S. Brisimi, T. Wang, W. Dai, and I. C. Paschalidis, "A joint sparse clustering and classification approach with applications to hospitalization prediction," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, 2016, pp. 4566–4571.

[18] T. S. Brisimi, T. Xu, T. Wang, W. Dai, W. G. Adams, and I. C. Paschalidis, "Predicting chronic disease hospitalizations from electronic health records: An interpretable classification approach," *Proceedings of the IEEE*, vol. 106, no. 4, pp. 690–707, 2018.

[19] T. S. Brisimi, T. Xu, T. Wang, W. Dai, and I. C. Paschalidis, "Predicting diabetes-related hospitalizations based on electronic health records," *Statistical Methods in Medical Research*, 2019.

[20] D. Bertsimas and R. Shioda, "Classification and regression via integer optimization," *Operations Research*, vol. 55, no. 2, pp. 252–271, 2007.

[21] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[22] S. Chatterjee, "Assumptionless consistency of the lasso," *arXiv preprint arXiv:1303.5817*, 2013.

[23] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *Journal of Machine Learning Research*, vol. 19, no. 13, 2018. [Online]. Available: http://jmlr.org/papers/v19/17-295.html

[24] T. L. Lai, C. Z. Wei *et al.*, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[25] Y. S. Chow and H. Teicher, *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2012.

[26] H.-F. Chen and L. Guo, *Identification and stochastic adaptive control*. Springer Science & Business Media, 2012.

[27] B. Nielsen, "Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms," *Econometric Theory*, vol. 21, no. 3, pp. 534–561, 2005.

[28] C. Wei *et al.*, "Adaptive prediction by least squares predictors in stochastic regression models with applications to time series," *The Annals of Statistics*, vol. 15, no. 4, pp. 1667–1682, 1987.

[29] Y. S. Chow, "Local convergence of martingales and the law of large numbers," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 552–558, 1965.