Fast max-affine regression via stochastic gradient descent

Seonho Kim and Kiryung Lee
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA
Email: kim.7604@osu.edu, kiryung@ece.osu.edu

***Abstract*—We consider regression of the max-affine model that combines multiple affine models via the max function. The max-affine model ubiquitously arises in applications such as multiclass classification and auction problems. It also generalizes the forward model in phase retrieval and rectifier linear unit activation function. We present a non-asymptotic convergence analysis of mini-batch stochastic gradient descent (SGD) for max-affine regression when the model is observed at random locations following the sub-Gaussianity with anti-concentration. Under these assumptions, a suitably initialized SGD converges linearly to the ground truth. Due to its low per-iteration cost, SGD converges faster than alternating minimization and gradient descent in run time. Our numerical results corroborate the presented theoretical results.**

I. Introduction

The *max-affine* model combines k affine models in the form of

$$y = \max_{j \in [k]} \left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j^\star \rangle + b_j^\star \right) \tag{1}$$

to produce a piecewise-linear mutivariate functions, where $\boldsymbol{x}$ and y respectively denote the covariate and the response, and $[k]$ denotes the set $\{1, \dots, k\}$. The max-affine model arises in applications in statistics, machine learning, economics, and engineering. For example, the max-affine model has been adopted for multiclass classification problems [1], [2] and simple auction problems [3], [4]. Moreover, the max-affine model also represents well-known models in signal processing and machine learning as special cases. For example, the instance of (1) for $k = 2$ with $b_1^\star = b_2^\star = 0$ and $\boldsymbol{\theta}_1^\star = -\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$ reduces to

$$y = |\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle| \tag{2}$$

which is the forward model in phase retrieval. Similarly, the rectified linear unit (ReLU)

$$y = \max(\langle \boldsymbol{x}, \boldsymbol{\theta}^\star \rangle, 0) \tag{3}$$

is written in the form of (1) for $k = 2$ with $\boldsymbol{\theta}_1^\star = \boldsymbol{0}$ and $\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}^\star$. Also note that (1) includes the affine cases of (2) and (3) by taking $b_1^\star$ and $b_2^\star$ as nonzero scalars (e.g. [5]).

The regression parameters in (1) can be estimated via a nonlinear least squares given by

$$\min_{\{\boldsymbol{\theta}_j, b_j\}_{j=1}^k} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \max_{j \in [k]} (\langle \boldsymbol{x}_i, \boldsymbol{\theta}_j \rangle + b_j) \right)^2 \tag{4}$$

where $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ denotes the set of observations. A minimizer to (4) is not uniquely determined since the estimator is invariant under the permutation of the indices for the component linear models. Furthermore, the nonlinearity in (1) with the max function makes the least squares in (4) non-convex. Therefore, it is challenging to estimate the ground-truth parameters even up to the inherent equivalence class.

To tackle the problem (4), alternating minimization has been widely used for max-affine regression [6]–[8]. Since the max-affine model in (1) combines the k affine models through the max function, the inputs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are naturally partitioned into disjoint subsets concerning the linear model that attains the maximum with each input. If the partition by the ground-truth linear models is known a priori, then the model parameters can be estimated via k decoupled linear least squares. The *least-squares partition algorithm* [6] iteratively refines the parameter estimate by alternating between partitioning and least squares. The partitioning step has been further improved in later studies [8], [9]. The resulting alternating algorithms have shown better empirical performance in computation and accuracy.

In a series of recent papers, Ghosh et al. proposed a spectral initialization method so that the least-squares partition algorithm from the spectral initialization can provide a theoretical performance guarantee [10]–[12]. They showed the linear convergence of the *alternating minimization* (AM) algorithm with the spectral initialization under the standard Gaussian covariate assumption in the presence of sub-Gaussian noise [12]. Moreover, they extended the convergence theory of the alternating minimization algorithm to a more general sub-Gaussian covariate model with the anti-concentration property [10], [11]. Their analysis outlines the sufficient number of samples required to accurately recover the parameters starting from a suitably initialized parameter.

We have developed an analogous theory for the first-order methods including the *gradient descent* (GD) and *stochastic gradient descent* (SGD) [13]. In this paper, we report a subset of these results on stochastic gradient descent in the noiseless case. A stochastic gradient descent method has been widely used to solve nonlinear least square problems [14]–[17]. In particular, as illustrated in Figure 1, SGD empirically outperforms GD and AM on the max-affine regression problem in the noiseless case.

Figure 1 compares AM, GD, and a mini-batch SGD on

SK and KL are supported in part by NSF CAREER Award CCF-1943201.

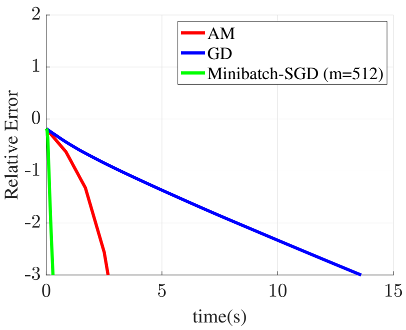

Fig. 1. Convergence of estimators for noise-free max-affine regression ($k = 3$, $d = 500$, and $n = 8,000$).

random 10 trials where the parameter vectors $\{\beta_j^\star\}_{j=1}^k$ are selected randomly from the unit sphere. We plot the median of relative errors versus the average run time. The relative error is measured up to the equivalence class given by the permutations of linear models, that is, the error is calculated as the minimum of $\log_{10}\left(\sum_{j=1}^k \|\widehat{\beta}_{\pi(j)} - \beta_j^\star\|_2^2 / \sum_{j=1}^k \|\beta_j^\star\|_2^2\right)$ over all possible permutation π over $[k]$, where $\{\widehat{\beta}_j\}_{j=1}^k$ denote the estimated parameters. While the theoretical analysis of AM and GD showed that both algorithms converge at least linearly to the ground-truth under comparable sample-complexity conditions, empirically, GD converges slower than AM in runtime. On the other hand, a mini-batch SGD converges much faster than AM in this experiment. Our main result derives a theoretical analysis of SGD that explains this empirical observation.

A. Main results

We perform the convergence analysis of the SGD estimator under the covariate model by [10]. They assumed that covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent copies of random $\boldsymbol{x} \in \mathbb{R}^d$ satisfying the following two properties.

Assumption 1.1 (Sub-Gaussianity): The covariate distribution satisfies

$$\|\langle \boldsymbol{v}, \boldsymbol{x}\rangle\|_{\psi_2} \le \eta, \quad \forall \boldsymbol{v} \in \mathbb{S}^{d-1}.$$

Assumption 1.2 (Anti-concentration): The covariate distribution satisfies

$$\sup_{w\in\mathbb{R}} \mathbb{P}((\langle \boldsymbol{v}, \boldsymbol{x}\rangle + w)^2 \le \epsilon) \le (\gamma\epsilon)^\zeta, \quad \forall \boldsymbol{v} \in \mathbb{S}^{d-1}.$$

The class of covariate distributions by Assumptions 1.1 and 1.2 generalize far beyond the standard independent and identically distributed Gaussian distribution. For example, the uniform and beta distributions satisfy Assumptions 1.1 and 1.2. Therefore, the result derived in this more general setting will apply to a wider range of applications.

This paper provides the first theoretical analysis of a mini-batch stochastic gradient descent estimator for max-affine regression under Assumptions 1.1 and 1.2. An informal version of the main result is stated below.

Theorem *Let $\boldsymbol{\beta}^\star \in \mathbb{R}^{k(d+1)}$ collect ground-truth parameters $(\boldsymbol{\theta}_j^\star, b_j^\star)_{j\in[k]}$. Within a neighborhood of $\boldsymbol{\beta}^\star$, a mini-batch stochastic gradient descent with $\widetilde{O}(C_{\boldsymbol{\beta}^\star} k^5 d)$ observations converges linearly to $\boldsymbol{\beta}^\star$ where $C_{\boldsymbol{\beta}^\star}$ is a constant that may implicitly depend on k through $\boldsymbol{\beta}^\star$, but are independent of d.*

In addition, we present numerical results demonstrating that stochastic gradient descent recovers the ground-truth parameters significantly faster from fewer observations than alternating minimization and gradient descent.

B. Organizations and Notations

The rest of the paper is organized as follows: Section II describes the gradient descent algorithm and introduces geometric parameters to state the main result in Section III. Section IV presents numerical results to demonstrate the empirical behavior of the gradient descent estimator. Finally, Section V concludes the paper with remarks and discussions on future directions.

Boldface lowercase letters denote column vectors, and boldface capital letters denote matrices. The concatenation of two column vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is denoted by $[\boldsymbol{a};\ \boldsymbol{b}]$. For column vector $\boldsymbol{a} \in \mathbb{R}^{d+1}$, a subvector of $\boldsymbol{a}$ with the first d entries will be denoted by $(\boldsymbol{a})_{1:d}$. Various norms are used throughout the paper. The spectral norm, Frobenius norm, Euclidean norm, and sub-Gaussian norm will be respectively denoted by $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_2$, and $\|\cdot\|_{\psi_2}$. Moreover, B_2^d and $\mathbb{S}^{d-1}$ will denote the d-dimensional unit ball and unit sphere with respect to the Euclidean norm. For two scalars q and p, we write $q \lesssim p$ if there exists an absolute constant $C > 0$ such that $q \le Cp$. We adopt the big-O notation so that $q \lesssim p$ is alternatively written as $q = O(p)$. With the tilde on the top of O, we ignore logarithmic factors. Lastly, the set $\{1, \ldots, n\}$ will be denoted by $[n]$ for $n \in \mathbb{N}$.

II. Problem Formulation

We formulate a mini-batch SGD algorithm for the least squares estimator. First, we rewrite the model in (1) as

$$y = \max_{j\in[k]} \langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star\rangle$$

where $\boldsymbol{\xi} := [\boldsymbol{x};\ 1] \in \mathbb{R}^{d+1}$ and $\boldsymbol{\beta}_j := [\boldsymbol{\theta}_j;\ b_j] \in \mathbb{R}^{d+1}$. Then the quadratic loss function is given by

$$\ell(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n \underbrace{\frac{1}{2}\left(y_i - \max_{j\in[k]}\langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j\rangle\right)^2}_{\ell_i(\boldsymbol{\beta})} \tag{5}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1;\ \ldots;\ \boldsymbol{\beta}_k] \in \mathbb{R}^{k(d+1)}$.

Let I_t be a set of m indices which are selected uniformly random with replacement from $[n]$ for $t \in \{0\} \cup \mathbb{N}$. Here the parameter m denotes the batch size. Then a mini-batch SGD with step size μ updates the estimate by

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \mu\frac{1}{m}\sum_{i\in I_t} \nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}^t), \quad \forall t \in \{0\} \cup \mathbb{N},$$

where $\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}^t)$ is explicitly computed as follows. A subgradient of each summand of the cost function in (5) with respect to the jth block $\boldsymbol{\beta}_j$ is written as

$$\nabla_{\boldsymbol{\beta}_j} \ell_i(\boldsymbol{\beta}) = \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}} \left(\max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle - y_i \right) \boldsymbol{\xi}_i, \quad \forall i \in [n], \tag{6}$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_k \subset \mathbb{R}^d$ are determined by $\boldsymbol{\beta}$ as

$$\begin{aligned} \mathcal{C}_j := \{ \boldsymbol{w} \in \mathbb{R}^d \ : \ & \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle > 0, \ \forall l < j \text{ and} \\ & \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j - \boldsymbol{\beta}_l \rangle \geq 0, \ \forall l > j \}. \end{aligned}$$

The set $\mathcal{C}_j$ contains all inputs maximizing the jth linear model.[1] Note that each $\mathcal{C}_j$ is determined by $k-1$ half spaces given by the pairwise difference of the jth linear model and the others. Then the gradient $\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$ is obtained by concatenating $\{\nabla_{\boldsymbol{\beta}_j} \ell_i(\boldsymbol{\beta})\}_{j=1}^k$ by

$$\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta}) = \sum_{j=1}^k \boldsymbol{e}_j \otimes \nabla_{\boldsymbol{\beta}_j} \ell_i(\boldsymbol{\beta}),$$

where $\boldsymbol{e}_j \in \mathbb{R}^k$ denotes the jth column of the k-by-k identity matrix $\boldsymbol{I}_k$ for $j \in [k]$. Moreover, $\ell_i(\boldsymbol{\beta})$ is differentiable except on a set of measure zero, with a slight abuse of terminology, $\nabla_{\boldsymbol{\beta}} \ell_i(\boldsymbol{\beta})$ in (6) is referred to as the "gradient".

III. Convergence analysis of mini-batch stochastic gradient descent

We present a convergence analysis of a mini-batch stochastic gradient descent estimator. The analysis depends on a set of geometric parameters of the ground-truth model. The first parameter $\pi_{\min}$ describes the minimum portion of observations corresponding to the linear model which achieved the maximum least frequently. It is formally defined as a lower bound on the probability measure on the smallest partition set, i.e.

$$\min_{j \in [k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star) \geq \pi_{\min}, \tag{7}$$

where $\mathcal{C}_1^\star, \ldots, \mathcal{C}_k^\star$ are polytopes determined by

$$\begin{aligned} \mathcal{C}_j^\star := \{ \boldsymbol{w} \in \mathbb{R}^d \ : \ & \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle > 0, \ \forall l < j \text{ and} \\ & \langle [\boldsymbol{w};\ 1], \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_l^\star \rangle \geq 0, \ \forall l > j \}. \end{aligned}$$

The next parameter κ quantifies the separation between all pairs of distinct linear models in (1) so that the pairwise distance on two distinct linear models satisfy

$$\min_{j' \neq j} \| (\boldsymbol{\beta}_j^\star)_{1:d} - (\boldsymbol{\beta}_{j'}^\star)_{1:d} \|_2 \geq \kappa. \tag{8}$$

Given the conditions in (7) and (8), we state our main result as the following theorem, which presents the local linear convergence of a mini-batch stochastic gradient descent estimator.

Theorem 3.1: Let $\delta \in (0, 1/e)$, $y_i = \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle$ for $i \in [n]$ with $\boldsymbol{\xi}_i = [\boldsymbol{x}_i;\ 1]$, and $\{\boldsymbol{x}_i\}_{i=1}^n$ being independent copies of $\boldsymbol{x} \in \mathbb{R}^d$ satisfying Assumptions 1.1 and 1.2.[2] Then there exist absolute constants $C, R > 0$ and $c, \nu \in (0, 1)$, for which the following statement holds with probability at least $1 - \delta$: If the initial estimate $\boldsymbol{\beta}^0$ belongs to a neighborhood of $\boldsymbol{\beta}^\star$ given by

$$\mathcal{N}(\boldsymbol{\beta}^\star) := \left\{ \boldsymbol{\beta} \in \mathbb{R}^{k(d+1)} \ : \ \max_{j \in [k]} \| \boldsymbol{\beta}_j - \boldsymbol{\beta}_j^\star \|_2 \leq \kappa \rho \right\}$$

with

$$\rho := \min \left(\frac{R \pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}}{4 k^{\zeta^{-1}}} \cdot \log^{-1/2} \left(\frac{k^{\zeta^{-1}}}{R \pi_{\min}^{\zeta^{-1}(1+\zeta^{-1})}} \right), \frac{1}{4} \right),$$

then for any $m < n$ and all $\boldsymbol{\beta}^\star$ satisfying (7) and (8), the sequence $(\boldsymbol{\beta}^t)_{t \in \mathbb{N}}$ by a mini-batch stochastic gradient descent method with step size $\mu = c \min(1, m/(d + \log(n/\delta)))$ satisfies

$$\begin{aligned} & \mathbb{E}_{I_t} \left\| \boldsymbol{\beta}^t - \boldsymbol{\beta}^\star \right\|_2 \leq \\ & \left(1 - \min \left(1, \frac{m}{d + \log(n/\delta)} \right) c \nu \right)^t \left\| \boldsymbol{\beta}_0 - \boldsymbol{\beta}^\star \right\|_2 \end{aligned} \tag{9}$$

for all $t \in \mathbb{N}$, provided that

$$n \geq C \pi_{\min}^{-4(1+\zeta^{-1})} k^4 \left(k d \log(n/d) + \log(1/\delta) \right).$$

There are a few remarks in order. First, Theorem 3.1 establishes a local convergence result of a mini-batch stochastic gradient descent. The basin of convergence is given as a neighborhood around the ground truth and the radius depends on k, κ, and $\pi_{\min}^{-1}$. Furthermore, the number of sufficient observations scales linearly in d and polynomial in $\pi_{\min}^{-1}$ and k. Note that the parameter κ becomes small when the ground-truth max-affine model involves similar affine models. Also the parameter $\pi_{\min}$ is small when there exist degenerate affine models which rarely attain the maximum. In either both of these cases, the regression problem becomes challenging. Theorem 3.1 explains how these parameters propagate to the requirements on the initialization and sufficient samples for the linear convergence.

To obtain the required initial estimate, one may use spectral initialization by [12, Algorithm 2, 3], which consists of dimensionality reduction followed by a grid search. They provided a performance guarantee of a spectral initialization scheme under the standard Gaussian covariate assumption [12, Theorems 2 and 3]. Therefore, the reduction of Theorem 3.1 to the Gaussian covariate case combined with [12, Theorems 2 and 3] provides a global convergence analysis of mini-batch stochastic gradient descent, which is comparable to that for alternating minimization [12]. Even in this case, the number of sufficient samples for the success of spectral initialization overwhelms that for the subsequent stochastic gradient descent

[1] In case of a tie when the multiple linear models attain the maximum for a given sample, we assign the sample to smallest index among the multiple indices. In the analysis with random $\boldsymbol{x}_i$s, the event of duplicate maximizing indices will happen with probability 0 for any absolutely continuous probability measure. Therefore, the choice of a tie-break rule will not affect the analysis.

[2] To simplify the presentation, we assume that the parameters η, ζ, γ in Assumptions 1.1 and 1.2 are fixed numerical constants in the statement and proof of Theorem 3.1. Therefore, any constant determined only by η, ζ, γ will be treated as a numerical constant.

step. Since multiple steps of their analysis critically depend on the Gaussianity, it remains an open question whether the result on the spectral initialization generalizes to the setting by Assumptions 1.1 and 1.2.

Second, the analysis of Theorem 3.1 focuses on the noiseless case and the linear convergence of SGD in the iteration index applies regardless of the batch size. The behavior of SGD in the noiseless max-affine regression is quite different from the existing analysis of SGD in general [18]. When the cost function in the form of $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$ is smooth and strongly convex, it has been shown that if $\boldsymbol{\beta}^\star$ is the minimizer of all summands $\{\ell_i(\boldsymbol{\beta})\}_{i=1}^{n}$, SGD converges linearly to $\boldsymbol{\beta}^\star$ [19, Theorem 2.1]. Although the cost function in the noiseless max-affine regression does not satisfy these properties, they hold locally near the ground truth, whence establishing the local linear convergence of SGD.

Third, Theorem 3.1 is directly comparable to the analogous result for alternating minimization under the same covariate model [11, Theorem 1]. The convergence parameter ν in (9) is larger than $3/4$ for alternating minimization. However, in our analysis, ν is smaller than $3/4$, particularly for large k and $\pi_{\min}^{-1}$, which would lead to slower convergence with respect to the iteration index. On the other hand, the per iteration cost of alternating minimization $O(knd^2)$ is higher than that of a mini-batch stochastic gradient descent $O(kmd)$ by a factor of dn/m. Therefore, as shown in Figure 1, for small batch sizes, the convergence speed of SGD in run time is faster than alternating minimization.

IV. Numerical results

We study the empirical performance of the mini-batch SGD estimator for max-affine regression. Further, we compare the mini-batch SGD to alternating minimization estimator [12] and full GD. Considered estimators start from the spectral initialization method by Ghosh et al. [12]. According to Theorem 3.1, the step size of the mini-batch SGD and GD are set to $1/2 \min(1, m/d)$ and $1/2$ respectively.

In all experiments, covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are generated as independent copies of a random vector following $\mathrm{Normal}(\mathbf{0}, \boldsymbol{I}_d)$.

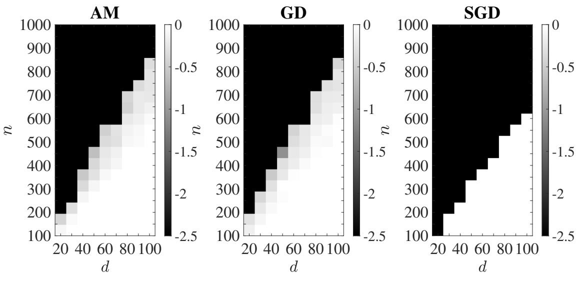

Fig. 2. Phase transition of estimation error per number of observations n and ambient dimension d in the noiseless case (number of linear models k and mini-batch size m are set to 3 and 64 respectively).

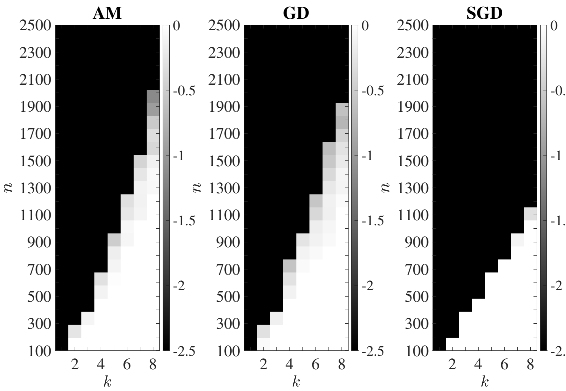

Fig. 3. Phase transition of estimation error per number of observations n and number of linear models k in the noiseless case (ambient dimension d and mini-batch size m are set to 50 and 64 respectively).

First, we observe the empirical phase transition of exact recovery in the noiseless case through Monte Carlo simulations. In this experiment, the ground-truth parameters $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_k^\star$ are generated as random mutually orthogonal vectors with $k < p$, and the offset terms are set to 0, i.e., $b_j^\star = 0$ for all $j \in [k]$. By the construction, the probabilities assigned to the maximizer set of each linear model become similar. In other words, $\pi_{\max}$ and $\pi_{\min}$ concentrate around $1/k$ where $\pi_{\max} := \max_{j \in [k]} \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)$. Furthermore, due to the orthogonality, the pairwise distance $\|\boldsymbol{\theta}_j^\star - \boldsymbol{\theta}_{j'}^\star\|_2 = \sqrt{2}$ for all $j \neq j' \in [k]$. Consequently, the sample complexity result for SGD by Theorem 3.1 simplifies to an easy-to-interpret expression $\widetilde{O}(k^{17}d)$ that involves only k and p. The result on alternating minimization [10] simplifies similarly. Figures 2 and 3 illustrate the empirical phase transition by the three estimators, where the median of normalized estimation error over 50 random trials is displayed. In these figures, the transition occurs when the sample size n becomes larger than a threshold that depends on the ambient dimension d and the number of linear models k. Figure 2 shows that the threshold for both estimators increases linearly with d for fixed k. This observation is consistent with the sample complexity by Theorem 3.1. A complementary view is presented in Figure 3 for varying k while d is fixed to 50. The threshold in Figure 3 for the SGD estimator is almost linear to k when p is fixed to 50. This rate is slower than the corresponding result in Theorem 3.1. A similar discrepancy between theoretical and empirical phase transitions has been observed for alternating minimization [10, Appendix L]. Figures 2 and 3 illustrates that a mini-batch stochastic gradient descent outperforms AM and GD. Since the inherent random noise in the gradient helps the estimator to escape the saddle points or local minima as studied in [20], [21], SGD recovers the parameter with fewer samples compared with the GD.

V. Conclusion

We have established a local convergence analysis of a mini-batch SGD for max-affine regression with noise-free observations. The covariate distribution is characterized by the sub-Gaussianity and anti-concentration, and generalizes beyond the standard Gaussian model. It has been shown that a mini-batch stochastic gradient descent estimator from a suitable initialization converges linearly to the ground truth. Due to a low per-iteration cost of SGD, overall, it provides faster convergence in run time than alternating minimization and GD.

References

[1] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of machine learning research*, vol. 2, no. Dec, pp. 265–292, 2001.

[2] A. Daniely, S. Sabato, and S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[3] J. Morgenstern and T. Roughgarden, "Learning simple auctions," in *Conference on Learning Theory*, pp. 1298–1318, PMLR, 2016.

[4] A. Rubinstein and S. M. Weinberg, "Simple mechanisms for a subadditive buyer and applications to revenue monotonicity," *ACM Transactions on Economics and Computation (TEAC)*, vol. 6, no. 3-4, pp. 1–25, 2018.

[5] B. Gao, Q. Sun, Y. Wang, and Z. Xu, "Phase retrieval from the magnitudes of affine linear measurements," *Advances in Applied Mathematics*, vol. 93, pp. 121–141, 2018.

[6] A. Magnani and S. P. Boyd, "Convex piecewise-linear fitting," *Optimization and Engineering*, vol. 10, no. 1, pp. 1–17, 2009.

[7] A. Toriello and J. P. Vielma, "Fitting piecewise linear continuous functions," *European Journal of Operational Research*, vol. 219, no. 1, pp. 86–95, 2012.

[8] L. A. Hannah and D. B. Dunson, "Multivariate convex regression with adaptive partitioning," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3261–3294, 2013.

[9] G. Balázs, "Convex regression: theory, practice, and applications," 2016.

[10] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression: Provable, tractable, and near-optimal statistical estimation," *arXiv preprint arXiv:1906.09255*, 2019.

[11] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression with universal parameter estimation for small-ball designs," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2706–2710, IEEE, 2020.

[12] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression: Parameter estimation for gaussian designs," *IEEE Transactions on Information Theory*, 2021.

[13] S. Kim and K. Lee, "Max-affine regression by first-order methods," *in preparation*.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[17] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[18] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa, "Tight analyses for non-smooth stochastic gradient descent," in *Conference on Learning Theory*, pp. 1579–1613, PMLR, 2019.

[19] D. Needell, R. Ward, and N. Srebro, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," *Advances in neural information processing systems*, vol. 27, 2014.

[20] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *International conference on machine learning*, pp. 1724–1732, PMLR, 2017.

[21] H. Daneshmand, J. Kohler, A. Lucchi, and T. Hofmann, "Escaping saddles with stochastic gradients," in *International Conference on Machine Learning*, pp. 1155–1164, PMLR, 2018.