Variable Selection for Max-Affine Regression via Sparse Gradient Descent

Haitham Kanj, Seonho Kim, and Kiryung Lee
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

***Abstract*—Sparse max-affine regression is introduced as a solution to variable selection for non-linear learning problems. A non-asymptotic local convergence analysis for Sparse Gradient Descent (Sp-GD) is presented when covariates are independent and identically distributed Gaussian random vectors. If at most s out of d covariates are actively contributing to the explanation of the target variable and the max-affine model combines a fixed number of linear models, then a suitably initialized Sp-GD linearly approaches the optimal solution and the true active covariates given $n = \mathcal{O}(s \log[(n \vee ed)/s])$ noise-free samples. Numerical Monte Carlo results corroborate these theoretical findings on the phase transition of exact parameter recovery.**

***Keywords*—** variable selection, non-linear regression, sparsity, max-affine

I. Introduction

This paper introduces a sparse max-affine model in an effort to solve variable selection for non-linear regression problems. The unique piece-wise linear geometry of the combination of linear models via the max function allows the modeling of non-linear dependencies between covariates and the target variable. Therefore, the max-affine model proved useful in several areas of signal processing and statistics such as clustering, classification, convex regression, and auction problems [1]. Simpler max-affine models have also been used in phase retrieval and neural networks with the Rectified-Linear Unit (ReLU) activation function family. In regression models where the number of available covariates is large, employing all variables for learning can lead to lower estimation accuracy especially when the covariates exhibit undesirable statistical properties (e.g. high covariate covariance or low covariate-target cross-covariance). The variable selection that refines covariates into a smaller subset of most contributing variables can mitigate such issues [2]. Indeed, there are applications in finance and economics in which targets are modeled as nonlinear mappings from a subset of many available covariates. One particular example is the high-dimensional non-linear wage equation from labor economics where variable selection has been shown to be of major significance [3], [4]. Therefore, it is beneficial to assume that a subset of covariates is sufficient to approximate real-life non-linear multivariate functions with the max-affine model. However, the extension of max-affine regression to sparse covariates has not yet been explored.

Variable selection is well explored in the literature when the covariate-target dependence is modeled as linear [2], [5]. The usual approach is to apply a regularizer to the loss function to force the sparsity of the weight vector (e.g. Lasso and basis pursuit). Other approaches to linear modeling propose feature selection algorithms via statistical analysis such as analyzing the conditional covariance of the covariates [6]. In practical applications, the covariate-target dependence often exhibits non-linearity not captured by linear models. There have been extensions of the variable selection to certain nonlinear models via basis change [7], [8]. In other words, the non-linear dependence is modeled as a combination of non-linear basis functions (e.g. spline basis [7]) as an attempt to reformulate the regression problem as linear. Then, variable selection in this context refers to limiting the number of newly defined covariates from the over-parameterized list. This approach is similar to the well-known order selection for polynomial regression [9]. Their model is restricted to decoupled nonlinearities applying separately to each covariate. However, the sparse max-affine model provides a way to study the *joint* nonlinear dependence on all covariates which has not been yet explored in the literature.

On the other hand, a line of research developed statistical analysis and efficient computational methods for max-affine regression [1], [10]–[12]. Ghosh et al. [13] established a non-asymptotic analysis of the alternating minimization (AM) algorithm under random covariates with independent stochastic noise. A subsequent work [14] proposed solving the max-affine regression using first-order methods including Gradient Descent (GD) and Stochastic Gradient Descent (SGD). It is shown that SGD converges faster than AM with comparable sample complexities.

This paper studies sparse max-affine regression via a computationally efficient Sparse Gradient Descent (Sp-GD) method. We present non-asymptotic convergence guarantees under the assumption that the covariates follow the standard Gaussian model, and only s out of a total of d covariates contribute to the target-covariate dependence. When the number of combined linear models k is fixed, Sp-GD linearly converges to the ground-truth regression parameters given $O(s \log(ed/s))$ noise-free data. Furthermore, the dependence on k remains the same compared to the previous result without sparsity [14]. All in all, this paper accomplishes a joint and efficient solution for simultaneous variable selection and max-affine regression. Additionally, Monte Carlo simulations are provided under different covariate models to corroborate the theoretical guarantees.

In this paper, lightface characters denote scalars, lowercase boldface characters denote column vectors, and uppercase boldface denotes matrices. We also adopt the symbols for the max and min operators in the lattice theory, i.e. $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ for $a, b \in \mathbb{R}$.

II. Sparse Max-affine Regression

Suppose that target variable $y \in \mathbb{R}$ is expressed by s-sparse covariates $\boldsymbol{x} \in \mathbb{R}^d$ via a piece-wise-linear multivariate function given by the *max-affine* model

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

where $[k] \triangleq \{1, \ldots, k\}$, and $\{(\boldsymbol{\theta}_j^\star, b_j^\star)\}_{j=1}^k$ denote the ground-truth coefficients of the $(d+1)$-dimensional hyper-planes. Note that (1) can be seen as a rank-k tropical polynomial under the max-plus algebra [15]. Moreover, the covariate sparsity is equivalent to the condition that the parameter vectors $\{\boldsymbol{\theta}_j^\star\}_{j=1}^k$ are jointly s-sparse. Sparse max-affine regression refers to the estimation of the parameters of the model in (1) from (noisy) dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.

This work was supported in part by NSF CAREER Award CCF-1943201.

III. Sparse Gradient Descent Algorithm

This section explains the projected gradient descent algorithm for sparse max-affine regression. To simplify notation, we reparametrize the max-affine model in (1) into a max-linear model

$$y = \max_{j\in[k]} \langle \boldsymbol{\xi}, \boldsymbol{\beta}_j^\star \rangle, \tag{2}$$

where $\boldsymbol{\beta}_j^\star \triangleq [\boldsymbol{\theta}_j^\star; b_j^\star]$ and $\boldsymbol{\xi} \triangleq [\boldsymbol{x}; 1]$ with the semicolon denoting vertical concatenation. Similarly, the concatenated covariate samples are given by $\boldsymbol{\xi}_i = [\boldsymbol{x}_i; 1]$ for $i \in [n] \triangleq \{1, \ldots, n\}$. Let $\boldsymbol{\beta} \triangleq [\boldsymbol{\beta}_1; \ldots; \boldsymbol{\beta}_k]$ denote the vertical concatenation of all k hyper-plane coefficient vectors $\{\boldsymbol{\beta}_j\}_{j=1}^k \subset \mathbb{R}^{d+1}$. We consider the estimator $\widehat{\boldsymbol{\beta}}$ that minimizes the Mean Squared Error (MSE) loss function

$$\ell(\boldsymbol{\beta}) \triangleq \frac{1}{2n} \sum_{i=1}^n \left(y_i - \max_{j\in[k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle \right)^2, \tag{3}$$

under the constraint that $\widehat{\boldsymbol{\beta}}$ belongs to Γ_s defined by

$$\Gamma_s \triangleq \left\{ [\boldsymbol{\alpha}_1; \ldots; \boldsymbol{\alpha}_k] \in \mathbb{R}^{k(d+1)} : \left\| \left(\sum_{j=1}^k [\boldsymbol{\alpha}_j]_l^2 \right)_{l=1}^d \right\|_0 \le s \right\}, \tag{4}$$

where $\|\cdot\|_0$ counts the number of nonzero entries and $[\boldsymbol{x}]_l$ denotes the l-th entry of the original vector. In other words, each element in Γ_s is rearranged into $(d+1)\times k$ matrix with at most s nonzero rows except the last row. That is, the last row is fixed to 1 as coefficients of the bias terms $\{b_j^\star\}_{j=1}^k$.

Sp-GD is a variant of the projected gradient descent algorithm to pursue the above estimators, where the gradient is substituted by the generalized gradient [16] and the step size varies across blocks adaptively with the iterates. We introduce a geometric object to describe the Sp-GD algorithm. Consider a partition of $\mathbb{R}^d$ determined by $\boldsymbol{\beta} = [\boldsymbol{\beta}_1; \ldots; \boldsymbol{\beta}_k] \in \mathbb{R}^{k(d+1)}$ as

$$\begin{aligned}
&\bigcup_{j=1}^k \mathcal{C}_j(\boldsymbol{\beta}) \cup \mathcal{V}(\boldsymbol{\beta}) = \mathbb{R}^d, \\
&\mathcal{C}_j(\boldsymbol{\beta}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_j \rangle > \max_{l\in[k]\setminus\{j\}} \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_l \rangle\}, \\
&\mathcal{V}(\boldsymbol{\beta}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^d \ : \ \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_j \rangle = \langle [\boldsymbol{x};\ 1], \boldsymbol{\beta}_l \rangle, \forall\ l \ne j \in [k]\}, \\
&\mathcal{C}_j(\boldsymbol{\beta}) \cap \mathcal{C}_l(\boldsymbol{\beta}) = \varnothing, \quad \mathcal{C}_j(\boldsymbol{\beta}) \cap \mathcal{V}(y) = \varnothing, \quad \forall\ l \ne j \in [k].
\end{aligned}$$

The partition sets $\{\mathcal{C}_j(\boldsymbol{\beta})\}_{j=1}^k$ are called as tropical open cells in the max-plus algebra [15] and $\mathcal{V}(\boldsymbol{\beta})$ denotes the tropical zero-set which corresponds to the boundary of the partition.

The algorithm starts by applying the block-wise gradient step

$$\boldsymbol{\alpha}_j^{t+1} = \boldsymbol{\beta}^t - \mu_j(\boldsymbol{\beta}^t) \nabla_{\boldsymbol{\beta}_j} \boldsymbol{\ell}\left(\boldsymbol{\beta}^t\right), \quad j \in [k],$$

where the partial generalized gradient with respect to the variables in the jth block $\boldsymbol{\beta}_j$ is written as

$$\nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j(\boldsymbol{\beta})\}} \left(\langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j \rangle - \max_{j\in[k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle \right) \boldsymbol{\xi}_i,$$

as presented in [14, Equation 7]. The step size used for updating the jth block is determined from the current iterate $\boldsymbol{\beta}^t$ as

$$\mu_j(\boldsymbol{\beta}^t) = \left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j(\boldsymbol{\beta}^t)\}} \right)^{-1}. \tag{5}$$

Next, the projection step is applied using the orthogonal projector $\boldsymbol{\Psi}_s$ onto the set Γ_s defined in (4), i.e.

$$\boldsymbol{\Psi}_s(\boldsymbol{\alpha}) = \operatorname*{argmin}_{\widetilde{\boldsymbol{\alpha}} \in \Gamma_s} \ \|\boldsymbol{\alpha} - \widetilde{\boldsymbol{\alpha}}\|_2^2.$$

Then $\boldsymbol{\Psi}_s$ enforces the joint s-sparsity assumption on the hyperplane parameters (excluding the last row which is always set to 1) by selecting the optimal minimizer from Γ_s under the ℓ_2-norm. Algorithm 1 outlines the projected gradient descent algorithm. This update rule applies recursively until the algorithm converges by satisfying $\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|_2 \le \gamma \|\boldsymbol{\beta}^t\|_2$ for a small numerical constant $\gamma > 0$.

Algorithm 1: Sparse Gradient Descent (Sp-GD)

Input: dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, sparsity level s, model rank k, step size μ, and initial estimate $\boldsymbol{\beta}^0$
while *stop condition is not satisfied* **do**
  for $(j = 1;\ j \le k;\ j = j+1)$ **do**
    $\boldsymbol{\alpha}_j^{t+1} \leftarrow \boldsymbol{\beta}_j^t - \mu_j^t(\boldsymbol{\beta}^t) \nabla_{\boldsymbol{\beta}_j} \boldsymbol{\ell}\left(\boldsymbol{\beta}^t\right)$
    $\boldsymbol{\beta}_j^{t+1} \leftarrow \boldsymbol{\Psi}_s(\boldsymbol{\alpha}^{t+1})$
  end
  $t \leftarrow t+1$
end
Output: final estimate $\widehat{\boldsymbol{\beta}} \leftarrow [\boldsymbol{\beta}_1^t; \ldots; \boldsymbol{\beta}_k^t]$

IV. Theoretical Results of Sp-GD

In this section, we present a local convergence analysis of Sp-GD under the standard Gaussian covariate assumption. One important performance-related parameter is $\pi_{\min} \in (0, \frac{1}{k})$ that controls the separation of the component affine models in (1) by imposing minimal probability measure on the maximizer set of each model, i.e.

$$\min_{j\in[k]} \mathbb{P}\left(\boldsymbol{x} \in \mathcal{C}_j(\boldsymbol{\beta}^\star)\right) \ge \pi_{\min}. \tag{6}$$

Another performance-related parameter $\kappa > 0$ enforces the discrepancy among the hyperplane coefficient vectors by

$$\min_{j'\ne j\in[k]} \|[\boldsymbol{\beta}_j^\star]_{1:d} - [\boldsymbol{\beta}_{j'}^\star]_{1:d}\|_2 \ge \kappa. \tag{7}$$

Now we state the main theorem under these two conditions.

Theorem IV.1. *Suppose that $\{\boldsymbol{x}_i\}_{i=1}^n$ are independent copies of the standard Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Let the certainty measure $\delta \in (0,1)$ be fixed. Then there exist absolute constants $C, R > 0$ and $\rho \in (0,1)$, for which the following statement holds for all s-sparse $\boldsymbol{\beta}^\star$ satisfying* (6) *and* (7) *with probability at least $1-\delta$. If the initial estimate $\boldsymbol{\beta}^0$ belongs to a neighborhood of $\boldsymbol{\beta}^\star$ given by*

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

with

$$\rho := \frac{R\pi_{\min}^{\frac{3}{4}}}{4k^2} \cdot \log^{-1/2}\left(\frac{k^2}{R\pi_{\min}^{\frac{3}{4}}} \right) \wedge \frac{1}{4},$$

and

$$n \ge C\big[(s\log(ed/s) + \log(k/\delta)) \vee s\log(n/s) \big] k^4 \pi_{\min}^{-12},$$

then the sequence $\left(\boldsymbol{\beta}^t\right)_{t\in\mathbb{N}}$ generated by Sp-GD with the step size in (5) *satisfies*

$$\sum_{j=1}^k \left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2 \le \rho^t \sum_{j=1}^k \left\|\boldsymbol{\beta}_j^0 - \boldsymbol{\beta}_j^\star\right\|_2, \quad \forall t \in \mathbb{N}.$$

Theorem IV.1 implies local linear convergence of Sp-GD in the noiseless case when the algorithm is properly initialized for the Gaussian covariate model. The sample complexity scales linearly with

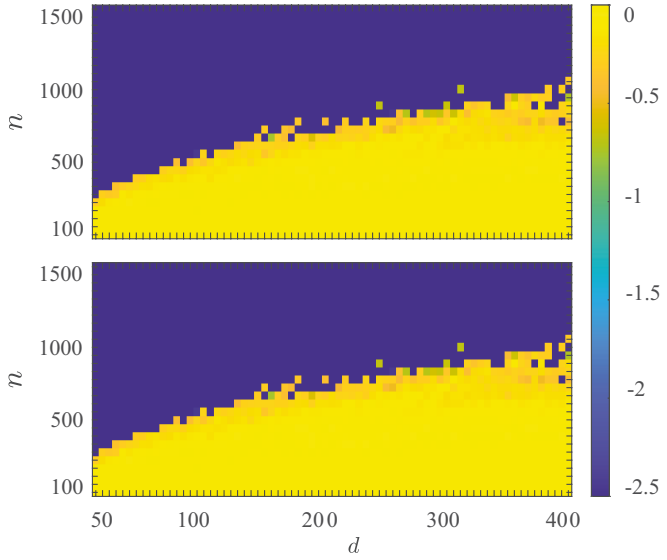

Fig. 1. Median of $E(\widehat{\boldsymbol{\beta}})$ for different (n,d) pairs using 50 Monte Carlo iterations for $k = 3$ and $s = 25$ with Gaussian (top) and Uniform (bottom) covariate distributions.

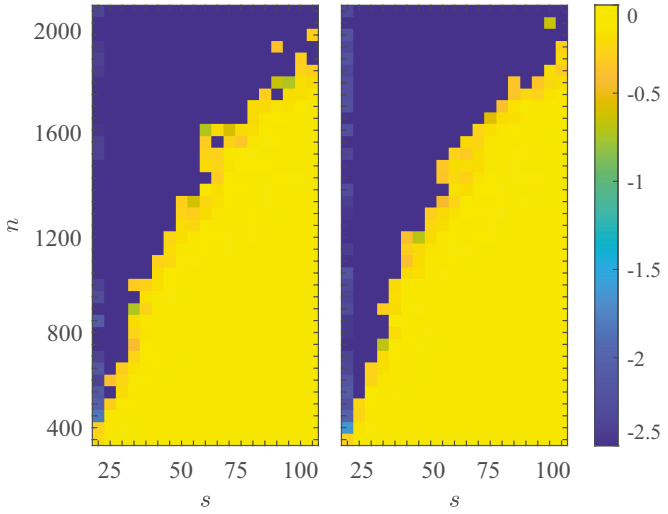

Fig. 2. Median of $E(\widehat{\boldsymbol{\beta}})$ for different (n,s) pairs using 50 Monte Carlo iterations for $k = 3$ and $d = 400$ with Gaussian (left) and Uniform (right) covariate distributions.

$s \ll d$ significantly improving analogous results without variable selection [13], [14]. Importantly, Sp-GD does not inflate the degree of dependence on the model rank number and dataset imbalance parameter and maintains the same order $k^4\pi_{\min}^{-12}$ as plain GD and SGD. Therefore, Sp-GD outperforms these algorithms regardless of the presence of sparsity in the max-affine model.

V. Numerical Results

This section presents the numerical results of the Sp-GD algorithm to corroborate the theoretical guarantees presented in IV. The estimation performance is evaluated via the median of the relative error between the true model coefficients $\boldsymbol{\beta}^\star \triangleq (\boldsymbol{\beta}_j^\star)_{j=1}^k$ and the estimated coefficients $\widehat{\boldsymbol{\beta}} \triangleq (\widehat{\boldsymbol{\beta}}_j)_{j=1}^k$. The relative error is defined via the optimal permutation of model indices as

$$E(\widehat{\boldsymbol{\beta}}) \triangleq \min_{\pi\in\mathrm{Perm}([k])} \log_{10}\left(\sum_{j=1}^k \|\widehat{\boldsymbol{\beta}}_{\pi(j)} - \boldsymbol{\beta}_j^\star\|_2^2 / \sum_{j=1}^k \|\boldsymbol{\beta}_j^\star\|_2^2\right).$$

Fig. 1 shows the empirical phase transition by Sp-GD per the total number of covariates d when the number of active covariates is fixed to $s = 25$ with $k = 3$. The phase transition occurs when n scales as a logarithmic function of d, corroborating the sample complexity in Theorem IV.1. Although Section IV presents theoretical guarantees under the Gaussian covariate assumption, Sp-GD provides a similar empirical phase transition under the uniform distribution of covariates as shown in the bottom of Fig. 1. Next, Fig. 2 shows the empirical phase transition by Sp-GD per the number of sparsely active covariates s when the total number of covariates is fixed to $d = 400$ with $k = 3$. This figure corroborates that the numerical complexity required for Sp-GD to work is indeed linearly dependent on s.

There is a spurious peak on the empirical phase transition boundary at $s = 10$ in both plots in Fig. 2. This phenomenon is due to the random generation of the ground-truth parameters in Monte Carlo simulations. If s is small, particularly, smaller than k, then the randomly generated parameter vectors $\{\boldsymbol{\beta}_j^\star\}_{j=1}^k$ have non-trivial correlations leading to a subset of similar linear component models in the max-affine model. It results in an imbalanced dataset, i.e. $\pi_{\min} < 1/k$. In contrast, for large s, randomly generated $\{\boldsymbol{\beta}_j^\star\}_{j=1}^k$ are almost orthogonal with high probability and $\pi_{\min} \approx 1/k$ providing a balanced dataset. This technical issue can be avoided if one explicitly controls generating the ground-truth parameter as an equiangular frame.

VI. Proof Sketch of Theorem IV.1

Considering the space limit of this conference paper, we only provide a sketch of the proof of Theorem IV.1. The proof is obtained by showing that each update in Sp-GD monotonically decreases the distance to the ground truth by a numerical constant factor, i.e.

$$\sum_{j=1}^k \left\|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\right\|_2 \le \rho \sum_{j=1}^k \left\|\boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star\right\|_2, \quad \forall t \in \mathbb{N}\cup\{0\}. \tag{8}$$

Define the support sets $\mathcal{S}^\star$ and $\mathcal{S}_j^t \subset [d+1]$ as the indices of the nonzero coefficients in $\boldsymbol{\beta}_j^\star$ and $\boldsymbol{\beta}_j^t$ with $\Theta_j^t \triangleq \mathcal{S}^\star \cup \mathcal{S}_j^t$, respectively. Let $\Theta_j^{t+1} \triangleq \mathcal{S}^\star \cup \mathcal{S}_j^{t+1}$ for all $j \in [k]$. For $\mathcal{S} \subset [d]$, the corresponding coordinate projection operators, $\boldsymbol{\Pi}_{\mathcal{S}} : \mathbb{R}^d \to \mathbb{R}^d$ and $\widetilde{\boldsymbol{\Pi}}_{\mathcal{S}} : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ are defined as

$$[\boldsymbol{\Pi}_{\mathcal{S}}\boldsymbol{x}]_j = \begin{cases} [\boldsymbol{x}]_j & \text{if } j \in S, \\ 0 & \text{otherwise,} \end{cases} \qquad \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}} = \begin{bmatrix} \boldsymbol{\Pi}_{\mathcal{S}} & 0 \\ 0 & 1 \end{bmatrix}.$$

Then each summand in the left-hand side of (8) is rewritten as

$$\begin{aligned} \|\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\|_2 &= \left\|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\beta}_j^\star\right)\right\|_2 \\ &\le \left\|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\alpha}_j^{t+1}\right)\right\|_2 + \|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\alpha}_j^{t+1} - \boldsymbol{\beta}_j^\star\right)\|_2 \\ &\le 2\|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\alpha}_j^{t+1} - \boldsymbol{\beta}_j^\star\right)\|_2 \\ &= 2\left\|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\beta}_j^t - \mu_j^t\nabla_{\boldsymbol{\beta}_j}\ell(\boldsymbol{\beta}^t) - \boldsymbol{\beta}_j^\star\right)\right\|_2, \end{aligned} \tag{9}$$

where the second inequality holds since $\|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\beta}_j^{t+1} - \boldsymbol{\alpha}_j^{t+1}\right)\|_2 \le \|\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\left(\boldsymbol{\alpha}_j^{t+1} - \boldsymbol{\beta}_j^\star\right)\|_2$, which follows from the fact that $\boldsymbol{\beta}^\star \in \Gamma_s$ and $\boldsymbol{\beta}^{t+1}$ is the projection of $\boldsymbol{\alpha}^{t+1}$ into Γ_s. We proceed with the following shorthand notations: $\boldsymbol{h}_j^t \triangleq \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_j^\star$, $\boldsymbol{v}_{jj'}^t \triangleq \boldsymbol{\beta}_j^t - \boldsymbol{\beta}_{j'}^t$, and $\boldsymbol{v}_{jj'}^\star \triangleq \boldsymbol{\beta}_j^\star - \boldsymbol{\beta}_{j'}^\star$. Due to the sparsity of the iterates and the ground-truth, we have $\boldsymbol{h}_j^t = \widetilde{\boldsymbol{\Pi}}_{\Theta^t}\boldsymbol{h}_j^t = \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}}\boldsymbol{h}_j^t + \widetilde{\boldsymbol{\Pi}}_{\Theta^t\setminus\Theta^{t+1}}\boldsymbol{h}_j^t$, and $\boldsymbol{v}_{jj'}^\star = \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star}\boldsymbol{v}_{jj'}^\star$.

For brevity, we introduce a shorthand notation $\mathcal{C}_j^t \triangleq \mathcal{C}_j(\boldsymbol{\beta}^t)$ for all $j \in [k]$. Then the partial gradient in (9) can be decomposed as

$$\begin{aligned}
&\widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \nabla_{\boldsymbol{\beta}_j} \ell(\boldsymbol{\beta}^t) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t\}} \left(\langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^t \rangle - \max_{j \in [k]} \langle \boldsymbol{\xi}_i, \boldsymbol{\beta}_j^\star \rangle \right) \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{\xi}_i \\
&= \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t\}} \langle \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{\xi}_i, \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t \rangle \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{\xi}_i}_{\boldsymbol{a}} \\
&\quad + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t\}} \langle \widetilde{\boldsymbol{\Pi}}_{\Theta^t \setminus \Theta^{t+1}} \boldsymbol{\xi}_i, \widetilde{\boldsymbol{\Pi}}_{\Theta^t \setminus \Theta^{t+1}} \boldsymbol{h}_j^t \rangle \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{\xi}_i}_{\boldsymbol{b}} \\
&\quad + \underbrace{\frac{1}{n} \sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t \cap \mathcal{C}_{j'}^\star\}} \langle \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star} \boldsymbol{\xi}_i, \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star} \boldsymbol{v}_{jj'}^\star \rangle \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{\xi}_i}_{\boldsymbol{c}} . \quad (10)
\end{aligned}$$

Plugging (10) into (9) yields

$$\frac{1}{2} \| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^{t+1} \|_2 \leq \| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t - \mu_j^t \boldsymbol{a} \|_2 + \mu_j^t \left(\| \boldsymbol{b} \|_2 + \| \boldsymbol{c} \|_2 \right). \quad (11)$$

We derive a probabilistic upper bound on each summand on the right-hand side of (11) by leveraging the following result. Let $\pi_j \triangleq \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j)$ and $\pi_j^\star \triangleq \mathbb{P}(\boldsymbol{x} \in \mathcal{C}_j^\star)$. Then similar to [14, Lemma 7.6] we can show that with high probability

$$\boldsymbol{\beta}^t \in \mathcal{N}(\boldsymbol{\beta}^\star) \implies (1-\eta) \leq \frac{\pi_j}{\pi_j^\star} \leq \left(\frac{1-\eta}{1-2\eta} \right) \quad (12)$$

for a small numerical constant $\eta > 0$. Furthermore, due to the Vapnik-Chervonenkis theory [17], the empirical measure $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}}$ concentrates around the expectation, π_j. Therefore, with high probability, we can have using (12)

$$\mu_j \triangleq \frac{1}{\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j\}}} \leq \frac{1}{\pi_j - \epsilon} \leq \frac{1}{(1-\eta)\pi_j^\star - \epsilon}.$$

With high probability, the first factor on the right-hand side is upper-bounded by

$$\frac{\| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \left(\boldsymbol{h}_j^t - \mu \boldsymbol{a} \right) \|_2}{\| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t \|_2} \leq \mu_j \epsilon, \quad \frac{\mu_j \| \boldsymbol{b} \|_2}{\| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t \|_2} \leq \mu_j \epsilon.$$

The vector $\boldsymbol{c}$ in the last term of (10) is factorized as $\boldsymbol{c} = \boldsymbol{E}_{\Theta^{t+1}} \boldsymbol{v}$, where $\boldsymbol{E}_{\Theta^{t+1}} \triangleq \left[\boldsymbol{\xi}_{1,\Theta^{t+1}}, \ldots, \boldsymbol{\xi}_{n,\Theta^{t+1}} \right]$ and

$$\boldsymbol{v} \triangleq \sum_{j' : j' \neq j} \begin{bmatrix} \mathbb{1}_{\{\boldsymbol{x}_1 \in \mathcal{C}_j^t \cap \mathcal{C}_{j'}^\star\}} \langle \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star} \boldsymbol{\xi}_1, \boldsymbol{v}_{jj'}^\star \rangle \\ \vdots \\ \mathbb{1}_{\{\boldsymbol{x}_n \in \mathcal{C}_j^t \cap \mathcal{C}_{j'}^\star\}} \langle \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star} \boldsymbol{\xi}_n, \boldsymbol{v}_{jj'}^\star \rangle \end{bmatrix}.$$

Moreover, by the restricted isometry property of a standard Gaussian matrix [18], we obtain

$$\| \boldsymbol{c} \|_2 \leq \frac{(1+\epsilon) \| \boldsymbol{v} \|_2}{\sqrt{n}}.$$

We can also obtain

$$\begin{aligned}
\frac{1}{n} \| \boldsymbol{v} \|_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j' : j' \neq j} \mathbb{1}_{\{\boldsymbol{x}_i \in \mathcal{C}_j^t \cap \mathcal{C}_{j'}^\star\}} \langle \widetilde{\boldsymbol{\Pi}}_{\mathcal{S}^\star} \boldsymbol{\xi}_i, \boldsymbol{v}_{jj'}^\star \rangle^2 \\
&\leq \underbrace{\frac{2}{5e} \left(\frac{\pi_{\min}}{16} \right)^3}_{\lambda} k^{-1} \sum_{j' : j' \neq j} \| \boldsymbol{v}_{jj'}^t - \boldsymbol{v}_{jj'}^\star \|_2^2 \\
&= \lambda k^{-1} \sum_{j' : j' \neq j} \| \boldsymbol{h}_j^t - \boldsymbol{h}_{j'}^t \|_2^2 \\
&\leq \lambda k^{-1} \sum_{j' : j' \neq j} \left(\| \boldsymbol{h}_j^t \|_2^2 + \| \boldsymbol{h}_{j'}^t \|_2^2 \right),
\end{aligned}$$

where the first inequality follows from invoking [14, Lemma 7.7] with $\zeta = 1/2$, $\gamma = e$, and d substituted by s along the union bound argument over all possible $\binom{d}{s}$ supports. Note that this substitution will inflate the error probability δ by $\log\left(\binom{n}{s}\right) \leq s \log\left(\frac{en}{s}\right)$. Therefore, this lemma holds for

$$n \geq C_2 k^4 \pi_{\min}^{-12} \left[s \log(ed/s) + \log(k/\delta) \right] \vee \left[s \log(n/s) \right],$$

for some absolute constant $C_2 > 0$. Since the ℓ_1 norm dominates the ℓ_2 norm, we can write

$$\frac{1}{\sqrt{n}} \| \boldsymbol{v} \|_2 \leq \sqrt{\frac{\lambda}{k}} \sum_{j' : j' \neq j} \left(\| \boldsymbol{h}_j^t \|_2 + \| \boldsymbol{h}_{j'}^t \|_2 \right).$$

Finally plugging the above upper bounds into (11) yields

$$\begin{aligned}
&\sum_{j=1}^{k} \| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^{t+1} \|_2 \\
&\leq \sum_{j=1}^{k} \left[\left(\frac{\epsilon}{(1-\eta)\pi_j^\star - \epsilon} \right) \left(\| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t \|_2 + \| \widetilde{\boldsymbol{\Pi}}_{\Theta^t \setminus \Theta^t} \boldsymbol{h}_j^t \|_2 \right) \right. \\
&\quad \left. + 4 \frac{\epsilon}{(1-\eta)\pi_j^\star - \epsilon} (k-1)(1+\epsilon) \sqrt{\frac{\lambda}{k}} \| \boldsymbol{h}_j^t \|_2 \right] \\
&\leq \frac{\epsilon}{(1-\eta)\pi_{min}^\star - \epsilon} \left(\sqrt{2} + 4(k-1)(1+\epsilon) \sqrt{\frac{\lambda}{k}} \right) \sum_{j=1}^{k} \| \boldsymbol{h}_j^t \|_2 \\
&\triangleq \rho \sum_{j=1}^{k} \| \boldsymbol{h}_j^t \|_2,
\end{aligned}$$

where the second inequality follows from

$$\| \widetilde{\boldsymbol{\Pi}}_{\Theta^{t+1}} \boldsymbol{h}_j^t \|_2 + \| \widetilde{\boldsymbol{\Pi}}_{\Theta^t \setminus \Theta^{t+1}} \boldsymbol{h}_j^t \|_2 \leq \sqrt{2} \| \widetilde{\boldsymbol{\Pi}}_{\Theta^t} \boldsymbol{h}_j^t \|_2,$$

and that trivially $\pi_j^\star \geq \pi_{\min}$ $\forall j \in [k]$. Finally, for a suitably chosen $\epsilon > 0$ one can have $\rho < 1$ which verifies the assertion in (8).

VII. Conclusion

This paper introduced sparse max-affine regression as a solution to variable selection for non-linear learning problems. A non-asymptotic local convergence analysis for Sparse Gradient Descent (Sp-GD) is presented when covariates are independent copies of the Gaussian random vector $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. To bypass the non-convexity of this optimization problem a suitable initialization method is employed. Then Sp-GD linearly approaches the optimal solution and the appropriate active covariates given $n = \mathcal{O}\left(s \log[(n \vee ed)/s]\right)$ noise-free samples when only s out of d covariates are actively contributing to the explanation of the target variable and the max-affine model combines a fixed number of linear models. Numerical Monte Carlo results corroborate these theoretical findings on the phase transition of exact parameter recovery for both Gaussian and uniform covariates. Finally, Sp-GD maintains the sample complexity dependence on the model rank number k and the dataset imbalance parameter $\pi_{\min}$ as plain GD and SGD. Therefore, regardless of sparsity in the max-affine model, Sp-GD always outperforms these algorithms.

REFERENCES

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

[2] L. Wasserman and K. Roeder, "High-dimensional variable selection," *The Annals of Statistics*, pp. 2178–2201, 2009.

[3] S. Singh, A. M. Haghighi, and S. Dalal, *Advanced Mathematical Techniques in Computational and Intelligent Systems*. CRC Press, 2023.

[4] C. Vance, "Marginal effects and significance testing with heckman's sample selection model: a methodological note," *Applied Economics Letters*, vol. 16, no. 14, pp. 1415–1419, 2009.

[5] E. I. George, "The variable selection problem," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1304–1308, 2000.

[6] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] L. Meier, S. Van de Geer, and P. Bühlmann, "High-dimensional additive modeling," *The Annals of Statistics*, pp. 3779–3821, 2009.

[8] V. Koltchinskii and M. Yuan, "Sparsity in multiple kernel learning," *The Annals of Statistics*, pp. 3660–3695, 2010.

[9] J. L. Peixoto, "Hierarchical variable selection in polynomial regression models," *The American Statistician*, vol. 41, no. 4, pp. 311–313, 1987.

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

[11] 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.

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

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

[14] S. Kim and K. Lee, "Max-affine regression via first-order methods," *arXiv preprint arXiv:2308.08070*, 2023. Accepted for publication in SIAM Journal on Mathematics of Data Science.

[15] P. Maragos, V. Charisopoulos, and E. Theodosis, "Tropical geometry and machine learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 728–755, 2021.

[16] J. Hiriart-Urruty, "New concepts in nondifferentiable programming," *Mémoires de la Société Mathématique de France*, vol. 60, pp. 57–85, 1979.

[17] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*, pp. 11–30, Springer, 2015.

[18] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.