Sparsity in Max-Plus Algebra And Applications in Multivariate Convex Regression

Nikos Tsilivis [1] Anastasios Tsiamis [2] Petros Maragos [1]

[1]School of ECE, National Technical University of Athens, Greece [2]ESE Department, SEAS, University of Pennsylvania, USA

ICASSP21 #: 3965

Convex Regression

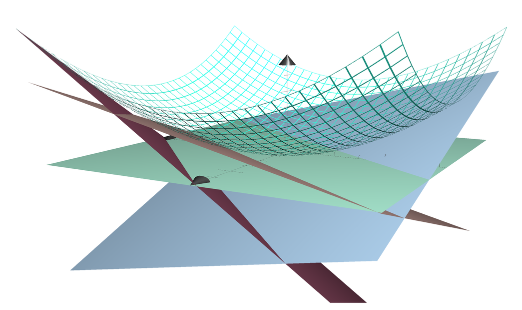
Data $(\mathbf{x}_i, y_i)$ from an unknown convex function f. How to estimate f?

Fundamental problem in optimization, signal processing, machine learning, and more!

Idea from Convex Analysis: Any convex function $\approx$ maximum of hyperplanes $\mathbf{a}_j^\intercal \mathbf{x} + \mathbf{b}_j$.

Estimation problem becomes a set of **nonlinear** equations over $(\mathbf{a}_j, b_j)_{j=1}^K$:

$$\begin{aligned} \max(\mathbf{a}_1^\intercal \mathbf{x}_1 + b_1, \ldots, \mathbf{a}_K^\intercal \mathbf{x}_1 + b_K) &= y_1 \\ \max(\mathbf{a}_1^\intercal \mathbf{x}_2 + b_1, \ldots, \mathbf{a}_K^\intercal \mathbf{x}_2 + b_K) &= y_2 \\ \ldots \end{aligned} \tag{1}$$



How to search for solutions in this problem? and how to keep the required number of parameters as small as possible?

Max-Plus Algebra

Originated from operations research and combinatorial optimization problems [1]. Has been applied successfully in areas such as Optimal Control, Nonlinear Signal and Image processing, Machine Learning.

Based on the tropical semiring $(\mathbb{R} \cup \{-\infty\}, \max, +)$, instead of the usual one $(\mathbb{R}, +, \times)$, max-plus algebra includes two key operations:

- Vector "addition": $\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} \bigvee \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} \max(x_1, y_1) \\ \max(x_2, y_2) \\ . \\ . \\ \max(x_n, y_n) \end{pmatrix}$.
- Vector "multiplication": $(x_1 \; x_2 \; . \; . \; x_n) \boxplus \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \max_{j=1}^n (x_j + y_j)$

Basen on the above, (1) can be written as:

$$\underbrace{\begin{pmatrix} \mathbf{a}_1^\intercal \mathbf{x}_1 & \mathbf{a}_2^\intercal \mathbf{x}_1 & \ldots & \mathbf{a}_K^\intercal \mathbf{x}_1 \\ . & . & . & . \\ . & . & . & . \\ \mathbf{a}_1^\intercal \mathbf{x}_m & \mathbf{a}_2^\intercal \mathbf{x}_m & \ldots & \mathbf{a}_K^\intercal \mathbf{x}_m \end{pmatrix}}_{\mathbf{A}} \boxplus \underbrace{\begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ b_K \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} y_1 \\ . \\ . \\ y_m \end{pmatrix}}_{\mathbf{y}} \tag{2}$$

Assuming known slopes $(\mathbf{a}_j)_{j=1}^K$ (either by numerically calculating the gradients of the data, or by discretizing an n-dimensional interval), this is a matrix max-plus equation over $\mathbf{b}$. Max-plus algebra equips us with tools to solve it *optimally*.

This approach works very well and yields a linear algorithm [2]. But can we be more **flexible**?

Sparsity and noise robustness

Key observation: If $b_j = -\infty$, then the whole $\mathbf{a}_j^\intercal \mathbf{x} + b_j$ hyperplane can be neglected, since the function never attains its value on it.

Thus, search for the solution of (2) that has the most $-\infty$ values, i.e the *sparsest*.

Definition 1. We call a vector $\mathbf{x}$ *sparse* if it contains many $-\infty$ elements.

Furthermore, we would like to account for the presence of noise in the data. That is, if $y_i = f(\mathbf{x}_i) + \epsilon$, then we expect equation (2) to hold only approximately. Thus, we are ultimately looking for **sparse approximate solutions** of (2).

Notice that relaxing the equality constraint promotes simpler models, as well.

But, how can we search for this kind of solutions?

Optimization in max-plus algebra

Definition 2. The *support set* of a vector is the set of indices of its values that are not equal to $-\infty$, that is: $\text{supp}(\mathbf{x}) = \{j \mid x_j \neq -\infty\}$.

Sparsity in max-plus algebra is computationally hard:

Theorem 1. Computing the sparsest solution of $\mathbf{A} \boxplus \mathbf{b} = \mathbf{y}$ is an NP-complete problem [3].

Note: It is essentially the minimum Set-Cover problem.

Based on the previous discussion on the convex regression problem, we formulate the following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{b}} |\text{supp}(\mathbf{b})| \\ &\text{s.t. } \|\mathbf{y} - \mathbf{A} \boxplus \mathbf{b}\|_p^p \leq \epsilon, \\ &\quad \mathbf{A} \boxplus \mathbf{b} \leq \mathbf{y}, \mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m. \end{aligned} \tag{3}$$

Notes:

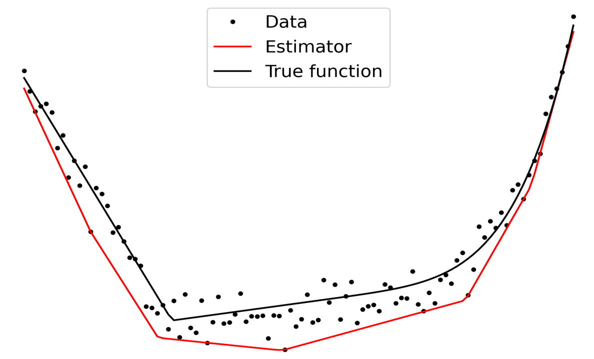
- we minimize the cardinality of the support set of the solution, while controlling the $\ell_p, p \leq \infty$, error.
- we also add a relaxation constraint $\mathbf{A} \boxplus \mathbf{b} \leq \mathbf{y}$, that restricts the approximation to happen from below (mainly for technical reasons).

Theorem 2. Problem (3) can be approximately solved in $\mathcal{O}(nm + n^2)$ time with a greedy algorithm.

Note: Submodular properties of the problem allows us to derive the approximation ratio of the algorithm, $\mathcal{O}(\log(m\|\mathbf{y}\|^p))$.

ℓ_∞ estimators

Approximating data from below might be problematic!



ℓ_∞ estimators - cont.

Can we drop the $\mathbf{A} \boxplus \mathbf{b} \leq \mathbf{y}$ constraint?

$$\begin{aligned} &\arg\min_{\mathbf{b}} |\text{supp}(\mathbf{b})| \\ &\text{s.t. } \|\mathbf{y} - \mathbf{A} \boxplus \mathbf{b}\|_\infty \leq \epsilon, \mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m. \end{aligned} \tag{4}$$

Proposition 1. *We can find a **locally optimal** solution of Problem (4) by solving Problem (3).*

*Technical details: Although a greedy algorithm for problem (4) might be arbitrarily bad, we can solve problem (3), add to the solution **half** of its ℓ_∞ error, and get a vector $\mathbf{b}^*$ that has the smallest ℓ_∞ error among all vectors with the same support set.*

Application to Multivariate Convex Regression

Based on the developed theory, we propose the following approach:

Input: Data $(\mathbf{x}_i, y_i) \in \mathbb{R}^{n+1}, i = 1, 2, \ldots, m$.

Model: $\max(\mathbf{a}_1^\intercal \mathbf{x}_i + b_1, \ldots, \mathbf{a}_K^\intercal \mathbf{x}_i + b_K) = y_i, \; i = 1, 2, \ldots, m$.

Step 1: Estimate slopes $(\mathbf{a}_j)_{j=1}^K$:

- Fixed values from an n-dimensional interval, or
- Numerical gradients of data.

Step 2: Solve Problem (3) (method called *Sparse Greatest Lower Estimate - SGLE*) or Problem (4) (called *Sparse Minimum Max Absolute Error -SMMAE estimate*), and calculate intercepts b_k.

Complexity: $\mathcal{O}(K^2 + K(n+1)m)$ or $\mathcal{O}(K^2 + K(n+2)m)$, respectively.

Output: A PWL convex approximation of the data with the approximately *minimum* number of affine regions needed for achieving the desired level of data fidelity.

Experiment on noisy paraboloid $z = x^2 + y^2 + \mathcal{N}(0, 1)$.

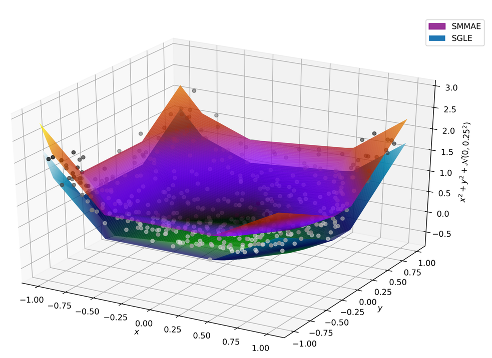


Figure 1:Approximation with 16 affine regions.

	SGLE		SMMAE		
(ϵ, p)	error$_{RMS}$	error$_\infty$	error$_{RMS}$	error$_\infty$	|supp|
$(300, 1)$	0.6681	1.5405	0.3506	0.7703	4
$(120, 2)$	**0.4899**	**1.1268**	0.2942	**0.5634**	31
$(150, 2)$	0.5465	1.1734	0.2729	0.5867	8
$(50, 5)$	0.5018	**1.1268**	0.2812	**0.5634**	23
$(10^8, 150)$	0.5560	**1.1268**	**0.2574**	**0.5634**	16

	GLE [2]		MMAE [2]	
K	error$_{RMS}$	error$_\infty$	error$_{RMS}$	error$_\infty$
10	0.6659	1.6022	0.3641	0.8011
25	0.5674	1.2779	0.3016	0.6389
100	0.5364	1.2828	0.3135	0.6414

Table 1:PWL approximations and their errors. K is the number of affine regions in the resulting tropical polynomial.

References

[1] R. Cuninghame-Green. *Minimax Algebra*. Springer-Verlag, 1979.

[2] P. Maragos and E. Theodosis. ``Multivariate Tropical Regression and Piecewise-Linear Surface Fitting''. In: *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 3822–3826.

[3] A. Tsiamis and P. Maragos. ``Sparsity in Max-plus Algebra''. In: *Discrete Events Dynamic Systems* 29 (2019), pp. 163–189.