Piecewise Fits of Functions or Data using Optimal Regression Trees

Berk Öztürk

MIT, Department of Aeronautics and Astronautics

Objectives

Given data from a continuous function $f(x),\ x \in \mathcal{R}^p$ in the form of $Y \in \mathcal{R}^n,\ X \in \mathcal{R}^{n \times p}$, I aim to approximate the underlying function with a set of piecewise linear (PWL), piecewise nonlinear (PWN) and piecewise convex (PWC) fits with trust regions.

Introduction

In engineering, design is often informed by input/output data from expensive simulations and/or testing. More than often, the output of interest is a nonlinear and nonconvex function of the input data. Nonlinear function estimation is an unsolved problem and an active area of research [1].

Trees, and especially optimal regression trees (ORTs) [2], present an exciting opportunity to partition and fit data in a form compatible with optimization. Of particular interest are non-convex inequalities, where a PWL/PWC approximation without trust regions would result in the epigraph of the approximation being convex. With ORTs however, we can obtain both the convex approximations and their trust regions reliably.

To do this, I leverage ideas from:

- holistic regression (HR), to augment my data with nonlinear functions of the features,
- linear regression (LR) and support vector machines (SVMs), to optimally select features over which to split data,
- and ORTs, to find a locally optimal splitting of the data as well as its approximation.

This kind of 'fitting' has a number of potential applications, from global optimization via mixed integer convex programming, to surrogate/low-order modeling for real-time simulation or control.

Data

I focus not on the way that the underlying function is sampled, but how to make the best use of the existing data. I demonstrate the effectiveness of the method on a variety of functions and data, including but not limited to:

- Known uni- and multi-variate functions of data.
- Airfoil simulation data that has known log-convex structure over the variables.

Method

The chosen approach has the following steps:

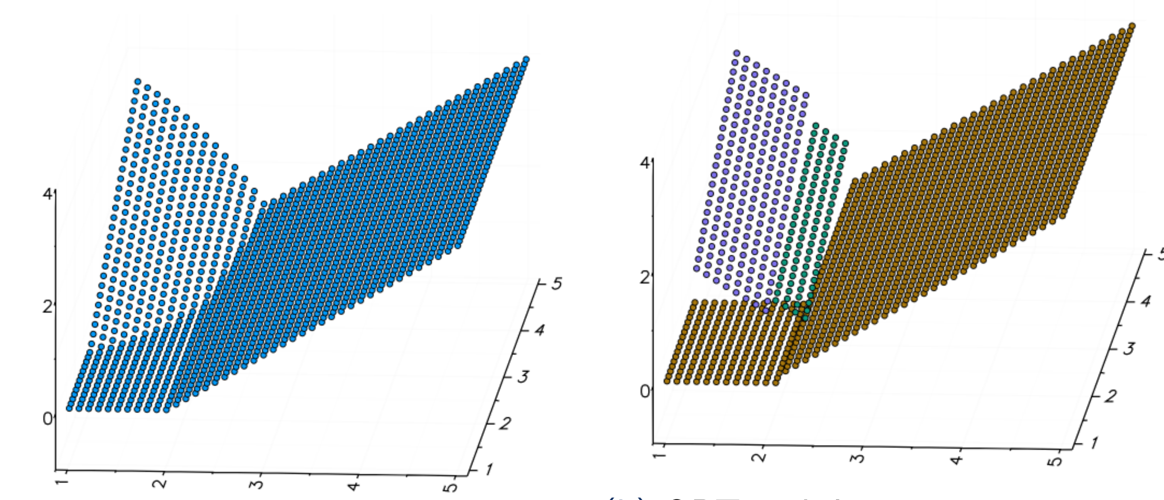
1. **Augment:** Choose a number of nonlinear functions of the parameters and add these to the data columns.
2. **Normalize and split:** Make sure all data is zero mean and variance of one, then split into training and test sets.
3. **Survey:** Use LR and/or SVMs to determine an appropriate level of sparsity in the split and regression features.
4. **Fit:** Fit ORTs over the data.
5. **Resolve:** Extract the trust regions and approximations of the function.

The approach is intrinsically amenable to being sparse in both the split and regressed features. Furthermore, it gives the designer the ability to tune the properties (eg. linearity, nonlinearity, convexity) of fits over the trust regions as desired, giving flexibility to the problem structure compared to other nonlinear fitting approaches.

Bivariate Demonstration - 1

I fit the following quasi-convex function (level sets of data are convex).

$$f(x) = \max(0,\ x_1 - \frac{1}{10}x_1x_2^2 - 2,\ x_2 - 2) \qquad (1)$$



(a) Original data

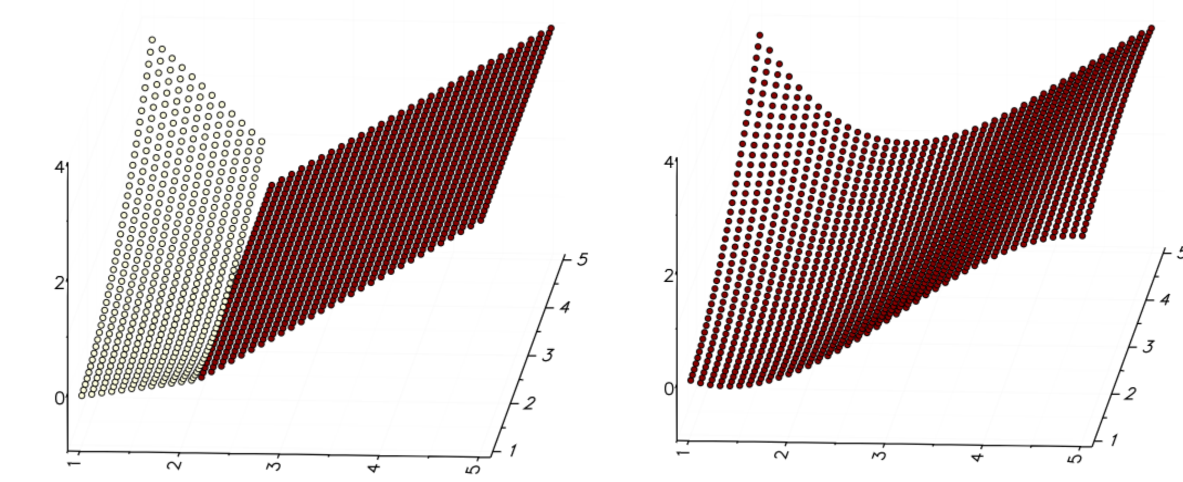
(b) ORT with linear regression, MSE 0.15.

Figure: Quasi-convex data predicted by piecewise linear regression.

Even with linear fits above, the MSE of prediction is low.

Bivariate Demonstration - 2

As I increase the complexity of my splitting and regression methods, I approximate the data using functions with different properties and accuracies, as shown below.



(a) ORT with second-order augmentations, MSE 0.22.

(b) ORT with third-order augmentations, MSE 0.13.

Figure: Additional complexity has potential to refine and impose different properties on the fits.

Main takeaway

Optimal Regression Trees built over a small number of nonlinear augmentations of data can reliably find good PWL/PWC/PWN approximations of underlying functions.

Univariate Demonstration - 1

$$f(x) = \max(-6x - 6,\ \frac{1}{2}x,\ \frac{1}{5}x^2 + \frac{1}{2}x) \qquad (2)$$

I approximate the above univariate, convex but discontinuous function, by augmenting the data $X = [x, x.^2, e^x, |x|]$, and fitting an optimal regression tree (ORT), only allow splits over the original feature x.

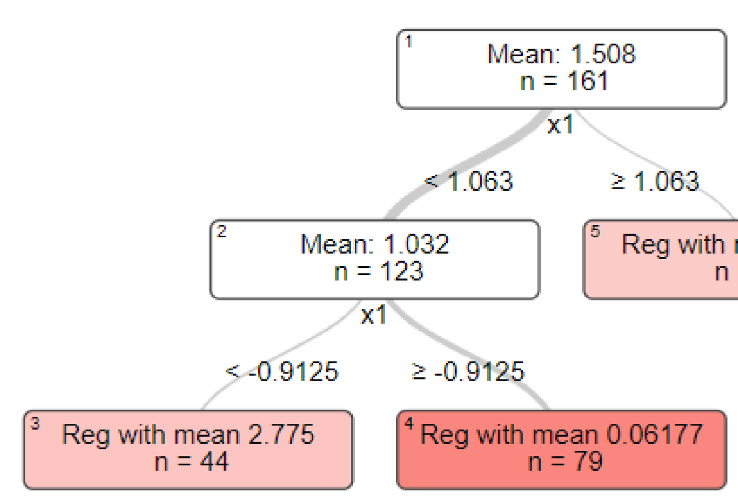


Figure: Resulting tree for fitting univariate function 2 over augmented data, $X = [x, x.^2, e^x, |x|]$.

Univariate Demonstration - 2

I see good matching between the data and predictions in the figure below, although the true optimum is not found. Different colors correspond to points in different leaves/PWC regions.

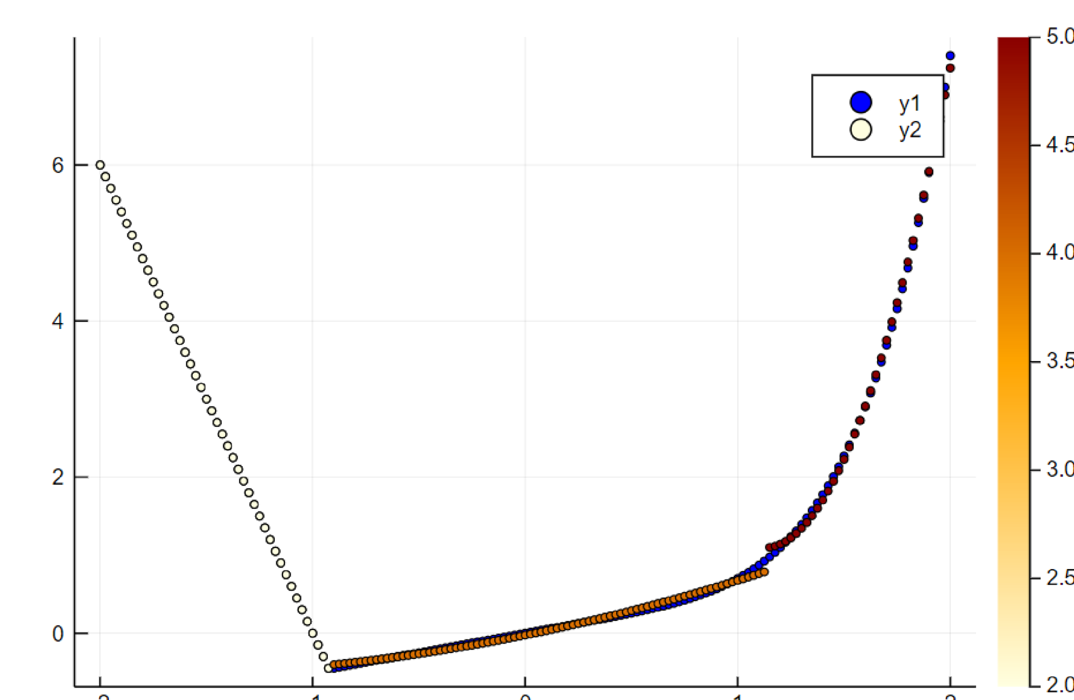


Figure: Quadratically augmented data recovers the underlying function with an RMS error $< 0.1\%$.

Conclusion and Future Work

ORTs over nonlinear augmentations of data represent a unique method to make different classes of approximations of functions or data. Future work will address challenges in the generality and scalability of this framework, accommodating more general nonlinear functions and higher dimensional data.

References

[1] Warren Hoburg, Philippe Kirschen, and Pieter Abbeel. Data fitting with geometric-programming-compatible softmax functions. *Optimization and Engineering*, 17(4):897–918, 2016.

[2] Dimitris Bertsimas and Jack Dunn. *Machine Learning Under a Modern Optimization Lens.* 2019.

[3] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Contact Information

- Web: https://convex.mit.edu/
- Email: bozturk@mit.edu