Optimization Methods for Nonparametric Convex Regression

Yunlang Zhu

December 18, 2023

Abstract

In this paper, we study the formulation of nonparametric convex regression as well as optimization methods for tackling this problem. Particularly, we discuss a general formulation of convex regression and then reformulate it to a computationally tractable quadratic programming problem. Furthermore, we present several recent papers that propose methods for solving convex regression with concerns on convergence guarantee, stability of solutions, and problem structure exploitation.

1 Introduction

Convex regression is a problem aimed at finding a convex function to best fit a finite number of observations. This technique raises interest in numerous fields since convexity (or concavity) is desired when fitting observations in several scenarios. For example, in financial engineering, the option pricing function is restricted to be convex under the no-arbitrage condition [1]. In addition to the desire for convexity, convex regression can help us apply prior knowledge to find a reasonable fitting function when the available datasets are too small in size to contain enough information to be exploited. For instance, experiments for the press hardening process in mechanical engineering are so expensive that it is hard to collect large datasets; therefore, prior knowledge that the hardness grows concavely with increasing temperature [3] can be introduced to compensate for this data shortage and to help us find a physically meaningful function with limited data.

To solve convex regression problems, there are usually two types of approaches, namely, parametric methods and nonparametric methods. In this paper, we will focus on optimization methods for nonparametric convex regression problems. Though we have witnessed advances in developing nonparametric methods, there are several challenges to be overcome, namely, the large-scale nature, the requirement for high accuracy, and the enormous complexity of the fitting functions obtained. In the remainder part of the paper, we will discuss the mathematical background of this topic in Section 2, as well as several recent works done in this field in Section 3.

2 Mathematical Background

To begin with, let us first recall the definitions of convex sets and convex functions. A set C is convex if for any two points $x, y \in C$, and for any $0 \leq \theta \leq 1$, we have

$$\theta x + (1-\theta)y \in C. \tag{1}$$

Moreover, if the domain of function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex set, and for any x, y in the domain and any $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \tag{2}$$

then f is a convex function. In addition, if f is differentiable, then it is convex if and only if

$$f(x) + \nabla f(x)^T(y-x) \leq f(y) \quad \forall x, y. \tag{3}$$

Now suppose we have n observations $\{(x_i, y_i)\} \in \mathbb{R}^d \times \mathbb{R}$ from our data, where $x_i \in \mathbb{R}^d$ is a vector for independent variables (features), and y_i is the corresponding response variable for x_i. Generally, we estimate the function of interest with a least squares estimator over all convex functions $\psi : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$y_i = \psi(x_i) + \epsilon \tag{4}$$

where ϵ is some noise with its expectation given the set of independent variables $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times d}$ being zero, that is, $\mathrm{E}[\epsilon|X] = 0$. Notice that we can perform concave regression as well, since the requirement that f is convex is equivalent to constraining $-f$ to be concave. The least squares estimator $\hat{\psi}$ is defined as

$$\hat{\psi} \in \arg\min_{\psi \in C} \sum_{i=1}^{n} (y_i - \psi(x_i))^2 \tag{5}$$

where $C := \{\psi : \mathbb{R}^d \to \mathbb{R} \mid \psi \text{ is a convex function}\}$. By deriving the least squares estimator, we are minimizing the squared error, also known as the l_2 loss function:

$$\mathrm{Loss}_{l_2} = \sum_{i=1}^{n} (y_i - \psi(x_i))^2. \tag{6}$$

Similarly, one can also perform optimization over the absolute error, also known as the l_1 loss function:

$$\mathrm{Loss}_{l_1} = \sum_{i=1}^{n} |y_i - \psi(x_i)|. \tag{7}$$

The optimization problem over functions in (5) appears to be infinitely dimensional and therefore intractable. However, it can be reformulated to an equivalent Quadratic Programming (QP) problem with $\theta_i = \hat{\psi}(x_i)$ for $i = 1, \ldots, n$; $\theta = (\theta_1, \ldots, \theta_n)^T$, and $Y = (y_1, \ldots, y_n)^T$ (this formulation is originally proposed by [6], which will not be discussed in this paper). The equivalent QP problem is

$$\begin{aligned}
\min_{\xi_1, \ldots, \xi_n; \theta} \quad & \frac{1}{2} ||Y - \theta||_2^2 && (8) \\
\text{s.t.} \quad & \theta_i + \xi_i^T (x_j - x_i) \leq \theta_j \quad \forall i, j, && (9) \\
& \theta \in \mathbb{R}^n, && (10) \\
& \xi_i \in \mathbb{R}^d \quad \forall i. && (11)
\end{aligned}$$

Notice that ξ_i in the formulation above (9) is the subgradient (or subderivative) at x_i. Subgradient is used here instead of derivative since the function might not be differentiable, e.g., the function can be piecewise linear. Mathematically, a vector $g \in \mathbb{R}^d$ is a subgradient of function ψ at a point x_i if

$$\psi(x_i) + g^T (x_j - x_i) \leq \psi(x_j) \quad \forall j. \tag{12}$$

Furthermore, the constraint (9) is used to ensure convexity.

3 Literature Review

In the literature, there are various methods for solving the convex regression problem (8) and its extensions. In this section, we review the relevant literature. Aiming at solving Problem (8), Mazumder et al. [5] propose an algorithmic framework based on augmented Lagrangian method and alternating direction method of multipliers (ADMM). First, they take the augmented Lagrangian of the original Problem (8). Then they employ a 3-block version of ADMM for the augmented Lagrangian (we refer our readers to [5] for full details). The cost for performing this algorithm is $O(\max\{n^2 d, nd^3\})$ with an additional $O(n^2 d^2 + nd^3)$ for the computation of matrix inverses for updating ξ_i. This algorithm leaves some questions, and two major ones are about convergence properties and method extension respectively. As a 3-block ADMM, the method has no convergence guarantee, since we can only secure convergence for 2-block ADMM instances. Though the paper proposes another convergence-guaranteed algorithm based on augmented Lagrangian method of multipliers, its computational cost is so high that it is needed to be initialized with a solution obtained from ADMM. On the other hand, the method cannot be easily extended for least absolute deviation convex regression, where the loss function is in l_1 form instead of l_2. In addition, the feasible set of Problem (8) can be unbounded. As a result, this may lead to instability, since different values of subgradients ξ_i can achieve the same cost function value.

In order to eliminate instability, Bertsimas and Mundru [2] add a regularization term to the objective function in Problem (8). The added regularization term leads to a strongly convex function, ensuring that ξ_i can only take some certain value for optimum. To extend this problem, [2] also proposes a formulation for l_1 convex regression. In order to guarantee convergence for solving the original Problem (8), which is not completely handled by [5], [2] presents a cutting-plane algorithm; this algorithm converges to an optimal solution within a finite number of iterations. The proposed method starts with an initial reduced master problem with $(n-1)$ constraints (the original problem has $n(n-1)$ constraints), which is assumed to be solved by Gurobi. Next, the solution found is verified with the entire set of constraints (we refer our readers to [2] for full details). Though [2] presents an approach to ensure convergence, this work applies an off-the-shelf solver, instead of developing a complete algorithm that exploits the special structure of this problem. Their work also discusses sparse convex regression, where the union of supports of ξ_i in each point x_i is a set whose cardinality is bounded by k, but that is out of our scope here.

To guarantee convergence while utilizing the structure of Problem (8), Lin et al. [4] propose a proximal augmented Lagrangian method (proxALM) with a semismooth Newton method for solving the proxALM subproblems. This paper also assumes additional shape constraints on subgradients ξ_i; the simply convex constraints are extended to more shape constraints, including monotone constraints, box constraints, and Lipschitz constraints. To solve the resulting problem, [4] develops a proxALM method. It is proximal since when updating θ and ξ, the algorithm calculates an approximate solution satisfying some stopping criterion. Particularly, updating θ, ξ is the most computationally intensive step. To deal with it, [4] designs a semismooth Newton method for this subproblem (we refer our readers to [4] for full details).

References

[1] Yacine Aït-Sahalia and Jefferson Duarte. Nonparametric Option Pricing Under Shape Restrictions. *Journal of Econometrics*, 116:9–47, 2003.

[2] Dimitris Bertsimas and Nishanth Mundru. Sparse Convex Regression. *INFORMS Journal on Computing*, 33(1):262–279, 2020.

[3] Martin von Kurnatowski, Jochen Schmid, Patrick Link, Rebekka Zache, Lukas Morand, Torsten Kraft, Ingo Schmidt, Jan Schwientek, and Anke Stoll. Compensating data shortages in manufacturing with monotonicity knowledge. *Algorithms*, 14(12):345, November 2021.

[4] Meixia Lin, Defeng Sun, and Kim-Chuan Toh. An Augmented Lagrangian Method with Constraint Generation for Shape-constrained Convex Regression Problems. *Mathematical Programming Computation*, 14:223–270, 2022.

[5] Rahul Mazumder, Arkopal Choudhary, Garud Iyengar, and Bodhisattva Sen. A Computational Framework for Multivariate Convex Regression and Its Variants. *Journal of the American Statistical Association (January 2018)*, pages 1–14, 2018.

[6] Emilio Seijo and Bodhisattva Sen. Nonparametric Least Squares Estimation of a Multivariate Convex Regression Function. *The Annals of Statistics*, 39(3):1633–1657, 2011.