Multivariate Distributionally Robust Convex Regression under Absolute Error Loss

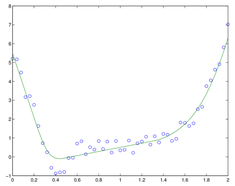
Jose Blanchet*, Peter W. Glynn*, Jun Yan† Zhengqing Zhou‡
* Stanford MS&E, † Stanford STATS, ‡ Stanford MATH
Email: {jose.blanchet, glynn, juyan65, zqzhou}@stanford.edu

Introduction

Consider a regression model of the form

$$Y_i = f_*(\mathbf{X}_i) + \mathcal{E}_i, \quad i = 1, 2, \cdots, n,$$

where the covariates $\mathbf{X}_i \in \mathbb{R}^d$, the response $Y_i \in \mathbb{R}$ and $\mathcal{E}_i$ are random noise with zero mean and finite variance.
We aim to non-parametrically estimate f_* under the **convex** shape constraint.



Why convex shape

Convexity is crucial in some settings, in the sense that a non-convex estimated function can create economically undesirable anomalies.

- For instance, consider the pricing of a European Call $C(K) = \mathbb{E}(S - K)^+$.
- Suppose we fit the price $\widehat{C}(K)$ that is **NOT** convex in K. This allows us to buy a contract at strike price $K + \varepsilon$ and one at $K - \varepsilon$, then sell two of the options at K.
- The total payoff at expiry is

  $P = (S-(K-\varepsilon))^+ + (S-(K+\varepsilon))^+ - 2(S-K)^+.$
- $P \geq 0$ always holds for all S (In particular, $P > 0$ for $K - \varepsilon < S < K + \varepsilon$), which leads to an **arbitrage**!

Applications

- In financial engineering, stock option prices usually have convexity restrictions.
- In economics, production functions and utility functions are often requried to be concave.
- Approximating an objective function for a convex optimization problem.
- Convex (concave) regression problems are also common in operations research and reinforcement learning.

Distributionally Robust Convex Regression

Q: What if we want an estimator **robust** to both adversarial perturbations in the empirical data and outliers?

$$\hat{f}^{\mathrm{DR}}_{n,\delta} := \underset{f \in \mathcal{F}}{\arg\min} \sup_{P \in \mathcal{P}(\mathbb{R}^{d+1}): D(P, P_n) \leq \delta} \mathbb{E}_P |Y - f(\mathbf{X})|,$$

where $\mathcal{F}$ is a suitable class of convex Lipschitz functions.

- Distributional robustness: By introducing the Wasserstien ball

  $$\{P : D(P, P_n) \leq \delta\},$$

  our estimator has performance guarantees under noisy inputs and small distributional shifts.
- Robust to outliers: Implement the L_1 loss function.

Tractable Formulation

Theorem: For any $\delta > 0$, we have

$$\inf_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}(\mathbb{R}^{d+1}): D(P, P_n) \leq \delta} \mathbb{E}_P |Y - f(\mathbf{X})| = \inf_{f \in \mathcal{F}} \left\{ \delta \|\nabla f\|_\infty + \frac{1}{n} \sum_{i=1}^n |Y_i - f(\mathbf{X}_i)| \right\}.$$

The inner maximization is solved in closed form resulting in a regularization with penalty on gradient.

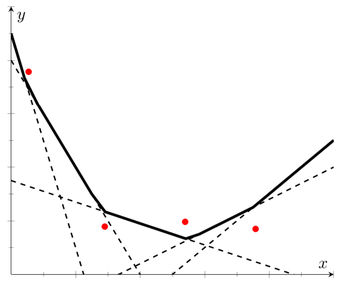
- To solve the DRCR, consider the following **finite dimensional** LP:

$$\begin{aligned} \min_{g_i, \xi_i} \quad & \frac{1}{n} \sum_{i=1}^n |Y_i - g_i| + \delta \max_{1 \leq i \leq n} \|\xi_i\|_\infty. \\ \text{s.t.} \quad & g_j \geq g_i + \langle \xi_i, X_j - X_i \rangle, \\ & 1 \leq i, j \leq n. \end{aligned}$$

- Let $(\widehat{g}_1, \widehat{\xi}_1), \cdots, (\widehat{g}_n, \widehat{\xi}_n)$ be any solution, then

$$\hat{f}^{\mathrm{DR}}_{n,\delta}(x) := \max_{1 \leq i \leq n} \left(\widehat{g}_i + \langle \widehat{\xi}_i, x - X_i \rangle \right).$$

- Example of our estimator ($d = 1$).



Statistical Guarantee

Theorem: Suppose $d > 4$, and the function f_* is convex with $\|\nabla f_*\|_\infty < \infty$.Under mild assumptions on the distribution of $\mathbf{X}$ and the noise $\mathcal{E}$, we can pick a δ_n of order $\widetilde{\Theta}(n^{-2/d})$ such that

$$l_1(\hat{f}^{\mathrm{DR}}_{n,\delta_n}, f_*) = \widetilde{\mathcal{O}}_P\left(n^{-1/d}\right).$$

Comparison to standard literature

	Existing Results	Our Result
Algorithm	QP, $\mathcal{O}(n^2)$ constraints	LP, $\mathcal{O}(n^2)$ constraints
$\mathbf{X}$	Bounded support	Light tail
Robustness	✗	✓
No apriori knowledge of f_*	✗	✓
Rate of Convergence	$\mathcal{O}\left(n^{-1/d}\right)$	$\widetilde{\mathcal{O}}\left(n^{-1/d}\right)$

Synthetic Datasets Analysis

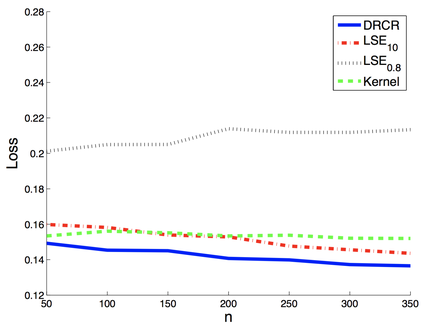
- Let $d = 5$, and $f_*(\mathbf{x}) = \|x\|_1$.
- Generate $\mathbf{X}_i$ i.i.d. from $N(\mathbf{0}, I_5)$, and let

  $$Y_i = f_*(\mathbf{X}_i) + \mathcal{E}_i,$$

  where $\mathcal{E}_i$ are sampled i.i.d. from $N(0, 0.04)$.
- To construct (DRCR) $\hat{f}^{\mathrm{DR}}_{n,\delta_n}$, we simply take

  $$\delta_n = n^{-2/5}.$$

- To compare, we consider the standard estimator (LSE$_c$) $\hat{f}^{\mathrm{LS}}_c$, which require an estimation $\|\nabla f_*\|_\infty \leq c$. We set $c = 10$ and 0.8, since in practise we may overestimate/underestimate the $\|\nabla f_*\|_\infty$.
- We also consider the standard kernel estimator.



Real Datasets Analysis

- We consider a public dataset from US Environmental Protection Agency, which consists of 600 air market data of California in 2019. The response was the amount of heat input with the covariates corresponding to the amounts of emissions of SO2, NOx, CO2 and the NOX rate.
- We implement three different approaches: DRCR, LSE and LR (linear regression).
- We repeat the experiment 10 times and then compare the average training l_1 loss and average test l_1 error.

Method	Training loss	Test error
DRCR	**0.1238**	**0.1294**
LSE	0.1485	0.1516
LR	0.1691	0.1692