Multivariate Distributionally Robust Convex Regression under Absolute Error Loss

Zhengqing Zhou

Department of Mathematics, Stanford University

Joint work with:

Jose H. Blanchet	Stanford MS&E
Peter W. Glynn	Stanford MS&E
Jun Yan	Stanford STATS

Problem Setting

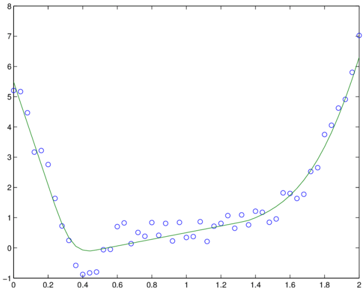
Stanford University

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.

Q: How to nonparametrically estimating f_* under the **convex** shape constraint?



Convex Shape Matters!

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, i.e., there exists K and $\varepsilon > 0$ such that

$$\widehat{C}(K+\varepsilon) + \widehat{C}(K-\varepsilon) < 2\widehat{C}(K).$$

- 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**!

Convex Regression

Convex regression problems arise in a variety of fields.

- 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.

- Model

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

- Estimator

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i))^2,$$

where $\mathcal{F}$ is a suitable convex function class.

- Measure of performance

$$l_p(f_*, \hat{f}_n) := \left[\frac{1}{n} \sum_{i=1}^{n} \Big(f_*(\mathbf{X}_i) - \hat{f}(\mathbf{X}_i) \Big)^p \right]^{1/p}.$$

where $p \geq 1$. In standard literature, l_2 error loss was used to measure the "distance" between $\hat{f}$ and f_*.

Recent Progress

Stanford University

- (Lim and Glynn'12, Seijo and Sen'11) Consistency of $\hat{f}_n$ for general d.
- (Guntuboyina and Sen'14) Suppose $d = 1$, the least square estimator $\hat{f}^{\mathrm{LS}}$ achieve the optimal rate

$$l_2(f_*, \hat{f}_n^{\mathrm{LS}}) = O_P(n^{-2/5}).$$

- (Lim'14) Suppose $d > 4$, $f_* : [0,1]^d \to \mathbb{R}$ and $\|\nabla f_*\|_\infty$ is bounded by C, then there is an estimator $\hat{f}_n$ (**depending on** C) such that

$$l_2(f_*, \hat{f}_n^{\mathrm{LS}}) = O_P(n^{-1/d}).$$

Our Model

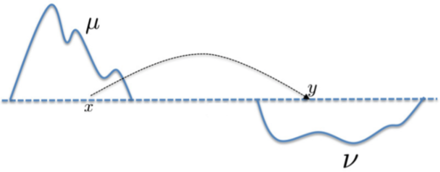
Stanford University

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

- Leverage the idea of Optimal Transport (Wasserstein distance).
- Let μ and ν be two distributions,

$$D_c(\mu, \nu) := \min \left\{ \mathbb{E}_\pi c(X, Y) : \pi \in \Pi(\mu, \nu) \right\},$$

where $c(\cdot, \cdot)$ is a cost function.



Our Model

Stanford University

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

Distributionally Robust Convex Regression (DRCR)

$(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ i.i.d.,

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

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

Remarks :

- 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

Stanford University

Theorem (Blanchet, Glynn, Yan and Z.'19)

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 penalty involves the norm of the gradient.
- This is still an infinite dimensional problem...

Algorithm

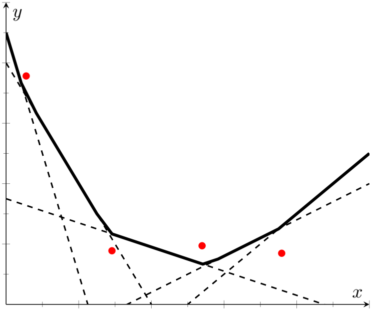
- To solve the DRCR, consider the following **finite dimensional** LP:

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

- Let $(\widehat{g}_1, \widehat{\xi}_1), \cdots, (\widehat{g}_n, \widehat{\xi}_n)$ be any solution of the above LP. Then

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

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



Statistical Guarantee

Stanford University

We focus on $d > 4$, and assume the function f_* is convex with $\|\nabla f_*\|_\infty < \infty$.

Theorem (Blanchet, Glynn, Yan and Z.'19)

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*

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

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	$\mathcal{O}\left(n^{-1/d}\right)$	$\widetilde{\mathcal{O}}\left(n^{-1/d}\right)$

Numerical Experiments

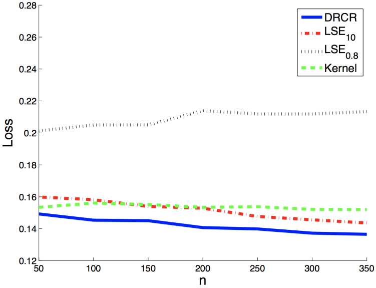
Stanford University

- Let $d = 5$, and $f_*(\mathbf{x}) = |x_1| + |x_2| + |x_3| + |x_4| + |x_5|$.
- Generate $\mathbf{X}_i$ i.i.d. from $N(\mathbf{0}, I_5)$, then generate Y_i by

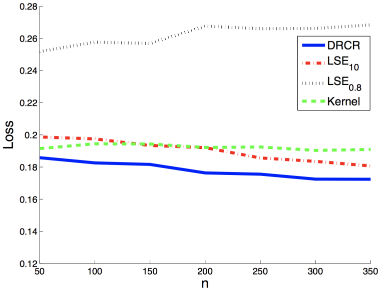
$$Y_i = f_*(\mathbf{X}_i) + \mathcal{E}_i,$$

  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 estimator (LSE$_c$) $\hat{f}^{\mathrm{LS}}_c$ in (Lim'14), 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 kernel estimator (typically not in convex shape), hyperparameters are chosen via cross-validation.
- We evaluate the performance in both l_1 and l_2 losses.

Numerical Experiments



(a) Light tail, l_1 loss



(b) Light tail, l_2 loss

- $\hat{f}^{\mathrm{DR}}_{n,\delta_n}$ outperforms $\hat{f}^{\mathrm{LS}}_{n,0.8}$, $\hat{f}^{\mathrm{LS}}_{n,10}$ and the kernel estimator in both l_1 and l_2 losses
- The performance of $\hat{f}^{\mathrm{LS}}_{n,c}$ is sensitive to the choice of the constant c, the a priori bound on $\|\nabla f_*\|_\infty$.

Conclusion

Stanford University

Summary

- We formulate and study the distributionally robust convex regression.
- Our estimator is designed to be robust to adversarial perturbations in the empirical data.
- Contrary to all of the existing results, our convergence rates hold without assuming compact domain, and with no apriori bounds on the underlying convex function or its gradient norm.

Future works

- Find the optimal rate of convergence and design an estimator to achieve it.
- Introduce distributional robustness to estimate other shape restricted functions, such as quasi-convex function, incresingly convex function, etc.

Thank you!