Robustness and Regularization:

Two sides of the same coin

(Joint work with Jose Blanchet and Yang Kang)

Karthyek Murthy
Columbia University
Jun 28, 2016

Introduction

- Richer data has tempted us to consider more elaborate models

$$\text{Elaborate models} \Longrightarrow \text{More factors / variables}$$

- Generalization has become a lot more challenging
- Regularization has been useful in avoiding overfitting

> Goal: A distributionally robust approach for improving generalization

Motivation for Distributionally robust optimization

- Want to solve the stochastic optimization problem

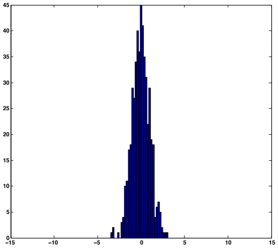
$$\min_{\beta} E\left[\text{Loss}(X, \beta)\right]$$

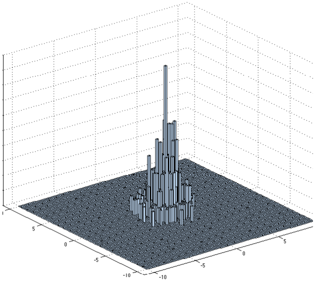
- Typically, we have access to the probability distribution of X only via its samples $\{X_1, \ldots, X_n\}$

- A common practice is to instead solve

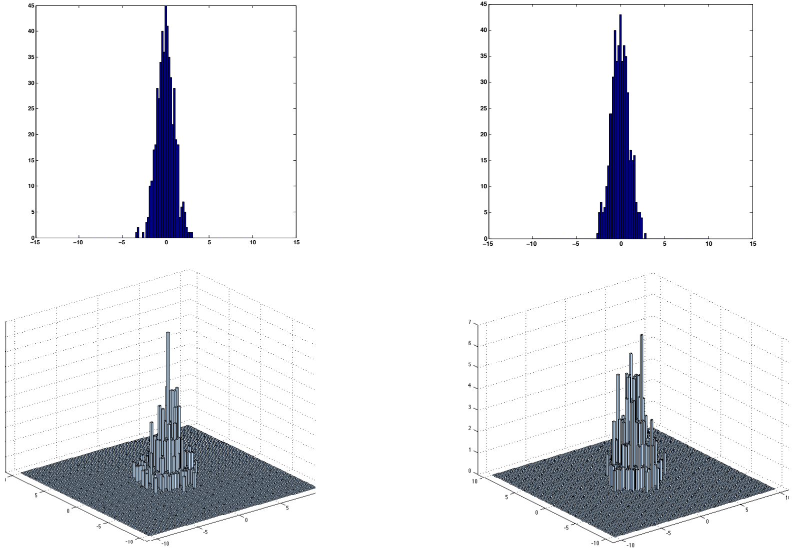
$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(X_i, \beta)$$

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(X_i, \beta) \quad \text{as a proxy for} \quad \min_{\beta} E\left[\text{Loss}(X, \beta)\right]$$





$$\min_\beta \frac{1}{n}\sum_{i=1}^{n} \text{Loss}(X_i, \beta) \quad \text{as a proxy for} \quad \min_\beta E\left[\text{Loss}(X, \beta)\right]$$



$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \text{Loss}(X_i, \beta) \quad \text{as a proxy for} \quad \min_{\beta} E\left[\text{Loss}(X, \beta)\right]$$

Learning

Natural to be thought as finding the "best" f such that

$$y_i = f(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n$$

$\mathbf{x}_i = (x_1, \ldots, x_d)$ is the vector of predictors

y_i is the corresponding response

[a]

[a] Image source: r-bloggers.com

Learning

Natural to be thought as finding the "best" f such that

$$y_i = f(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n$$

Empirical loss/risk minimization (ERM):

$$\frac{1}{n}\sum_{i=1}^{n} \text{Loss}\big(f(\mathbf{x}_i), y_i\big)$$

 [a]

[a] Image source: r-bloggers.com

Learning

Natural to be thought as finding the "best" f such that

$$y_i = f(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n$$

Empirical loss/risk minimization (ERM):

$$\frac{1}{n}\sum_{i=1}^{n} \text{Loss}\big(f(\mathbf{x}_i), y_i\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \big(y_i - f(\mathbf{x}_i)^2\big)$$

[a]

[a] Image source: r-bloggers.com

Learning

Natural to be thought as finding the "best" f such that

$$y_i = f(\mathbf{x}_i) + e_i, \qquad i = 1, \ldots, n$$

 [a]

[a] Image source: r-bloggers.com

Not enough

Find an f that fits well over "future" values as well

Generalization

Think of data $(\mathbf{x}_1, y_1), \dots (\mathbf{x}_n, y_n)$ as samples from a probability distribution P

Then "future values" can also be interpreted as samples from P

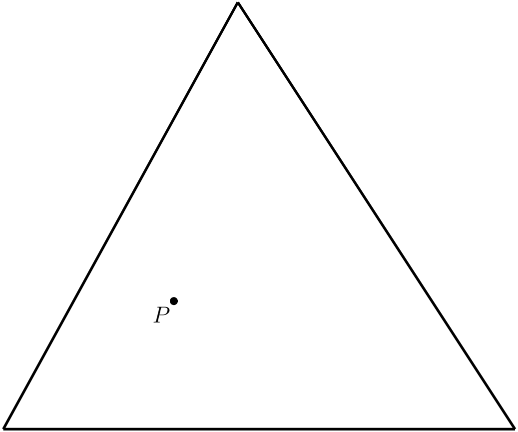
Generalization

Think of data $(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_n, y_n)$ as samples from a probability distribution P

Then "future values" can also be interpreted as samples from P

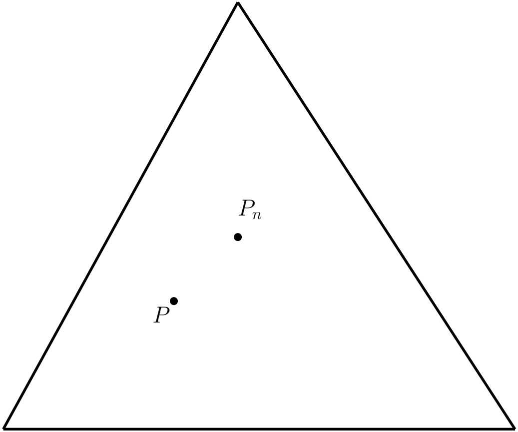
$$\min_f \frac{1}{n} \sum_{i=1}^{n} \text{Loss}\big(f(\mathbf{x}_i), y_i\big) \quad \longmapsto \quad \min_f E_P \left[\text{Loss}\big(f(X), Y\big)\right]$$

However, the access to P is still via samples, $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(\mathbf{x}_i, y_i)}$



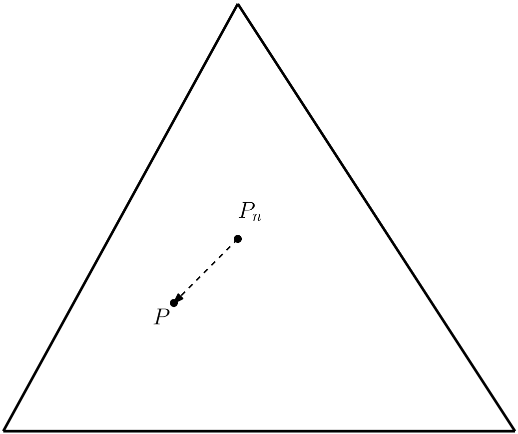
Want to solve $\min_{f \in \mathcal{F}} E_P \left[\text{Loss}\big(f(X), Y\big) \right]$

P unknown



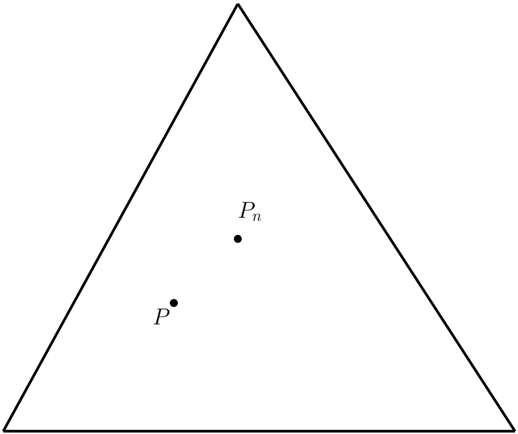
Know how to solve $\min_{f \in \mathcal{F}} E_{P_n} \left[\text{Loss}\big(f(X), Y\big) \right]$

Access to P via training samples P_n



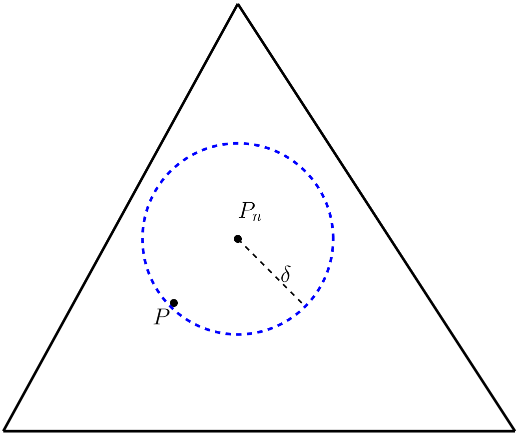
More and more samples give better approximation to P,

however, the quality of this approximation depends on dim



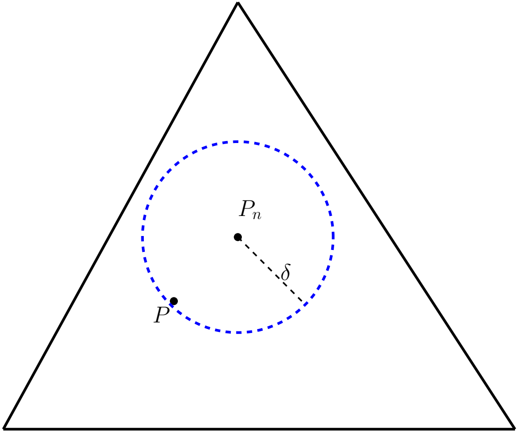
We are provided with only limited training data (n samples)

Sometimes, to an extent that even $n <$ dim of the parameter of interest.



Instead of finding the best fit with respect to P_n,

why not find a fit that works over all Q such that $D(Q, P_n) \leq \delta$



Formally,

$$\min_{f \in \mathcal{F}} \max_{Q : D(Q, P_n) \leq \delta} E_Q \left[\text{Loss}\big(f(X), Y \big) \right]$$

DR Regression:

$$\min_{f \in \mathcal{F}} \max_{Q: D(Q, P_n) \leq \delta} E_Q \left[\text{Loss}\big(f(X), Y\big) \right]$$

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q : D(Q, P_n) \leq \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

I. Are these DR regression problems solvable?
   - If so, how do they compare with known methods for improving generalization?

II. How to beat the curse of dimensionality while choosing δ?
   - Robust Wasserstein profile function

III. Does the framework scale?
   - Support vector machines
   - Logistic regression
   - General sample average approximation

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

How to quantify the distance $D(P, Q)$?

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

How to quantify the distance $D(P, Q)$?

Ans:

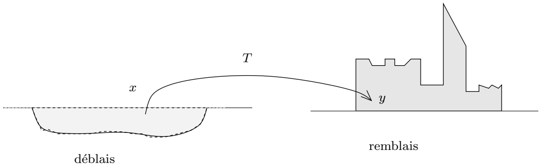
Let (U, V) be two random variables such that $U \sim P$ and $V \sim Q$.

Let us call a joint distribution (U, V) as π. Then

$$D(P, Q) = \inf_{\pi} E_\pi \| U - V \|$$

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$



[1]

How to quantify the distance $D(P, Q)$?

Ans:

Let (U, V) be two random variables such that $U \sim P$ and $V \sim Q$.

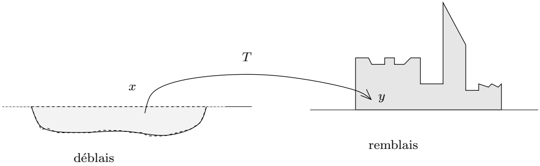
Let us call a joint distribution (U, V) as π. Then

$$D(P, Q) = \inf_{\pi} E_{\pi} \| U - V \|$$

[1] Image from the book Optimal Transport: Old and New by Cédric Villani

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$



How to quantify the distance $D(P, Q)$?

Ans:

Let (U, V) be two random variables such that $U \sim P$ and $V \sim Q$.

Let us call a joint distribution (U, V) as π. Then

$$D_c(P, Q) = \inf_{\pi} E_{\pi} \left[c(U, V) \right]$$

The metric D_c is called optimal transport metric.

When $c(u, v) = \|u - v\|^p$, $D_c^{1/p}$ is the p^{th} order Wasserstein distance

DR Linear Regression:

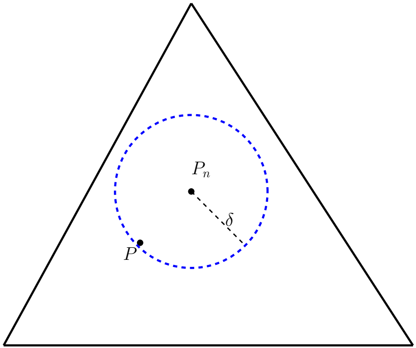
$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \leq \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

Next, how do we choose δ?

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

Next, how do we choose δ?



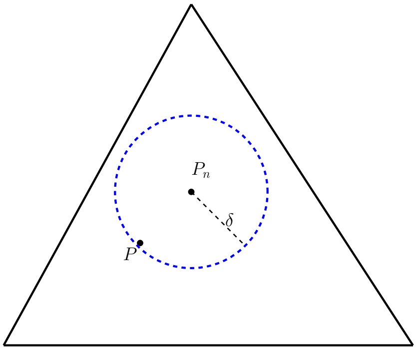
See Fournier and Guillin (2015), Lee and Mehrotra (2013), Shafieezadeh-Abadeh, Esfahani and Kuhn (2015)

DR Linear Regression:

$$\min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\leq\delta} E_Q\left[\left(Y-\beta^T X\right)^2\right]$$

The object of interest β_* satisfies:

$$E_P\left[\left(Y-\beta_*^T X\right)X\right]=0$$

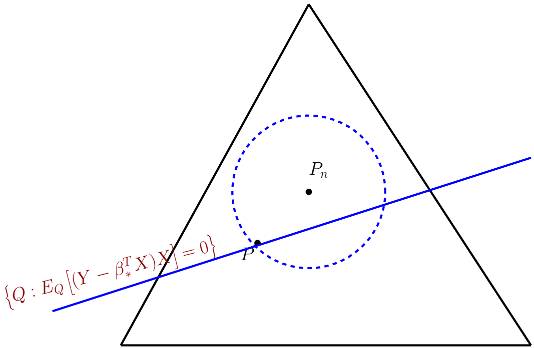


DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

The object of interest β_* satisfies:

$$E_P \left[\left(Y - \beta_*^T X \right) X \right] = 0$$

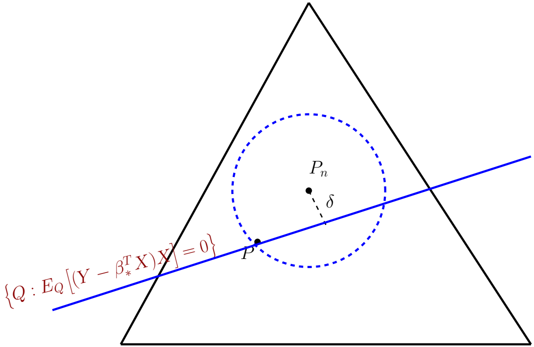


DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \leq \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

The object of interest β_* satisfies:

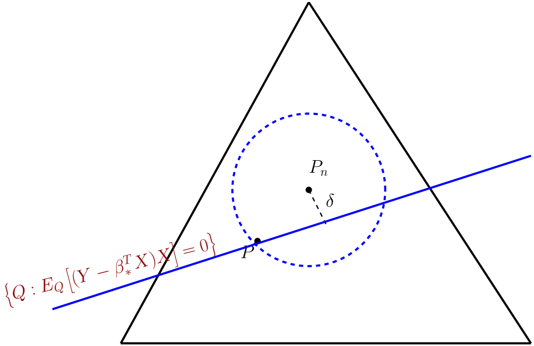
$$E_P\left[\left(Y - \beta_*^T X\right) X\right] = 0$$



$$R_n(\beta_*) = \min\left\{D_c\left(Q, P_n\right) : E_Q\left[\left(Y - \beta_*^T X\right) X\right] = 0\right\}$$

DR Linear Regression:

$$\min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\le\delta} E_Q\left[\left(Y-\beta^T X\right)^2\right]$$



Theorem 1
[Blanchet, Kang & M]
If $Y = \beta_*^T X + \epsilon$,

$$nR_n(\beta_*) \xrightarrow{D} \mathcal{L}$$

$$R_n(\beta_*) = \min\left\{D_c(Q,P_n) : E_Q\left[\left(Y-\beta_*^T X\right)X\right] = 0\right\}$$

DR Linear Regression:

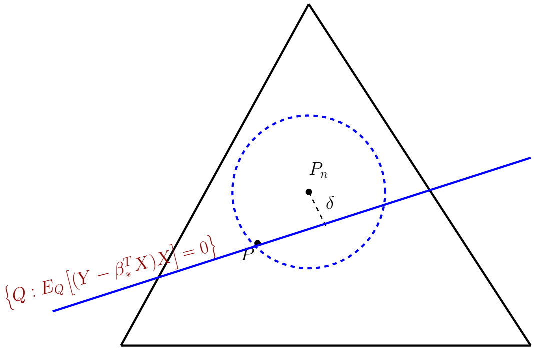
$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

Theorem 1

[Blanchet, Kang & M]

If $Y = \beta_*^T X + \epsilon$,

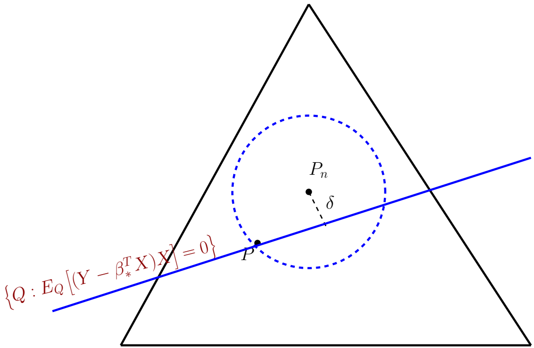
$$nR_n(\beta_*) \xrightarrow{D} \mathcal{L}$$



Choose $\delta = \dfrac{\eta}{n}$ where η is such that $P\{\mathcal{L} \leq \eta\} \geq 0.95$

DR Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$



Theorem 1

[Blanchet, Kang & M]

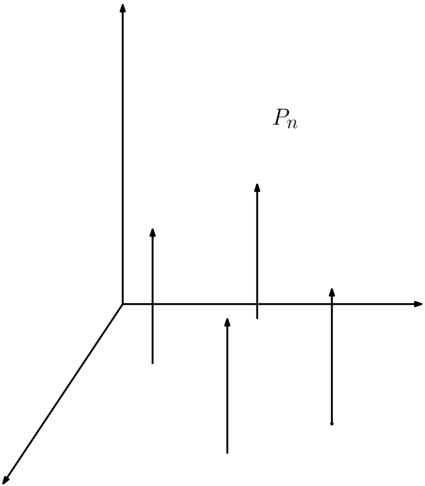
If $Y = \beta_*^T X + \epsilon$,

$$nR_n(\beta_*) \xrightarrow{D} \mathcal{L}$$

Choose $\delta = \frac{\eta_\alpha}{n}$ where η_α is such that $P\{\mathcal{L} \le \eta_\alpha\} \ge 1 - \alpha$.

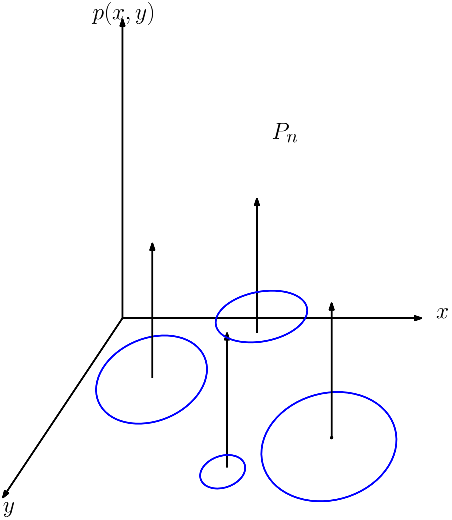
Robust Wasserstein profile function:

$$R_n(\beta) = \min\left\{D_c\left(Q, P_n\right) : E_Q\left[\left(Y - \beta^T X\right)X\right] = 0\right\}$$



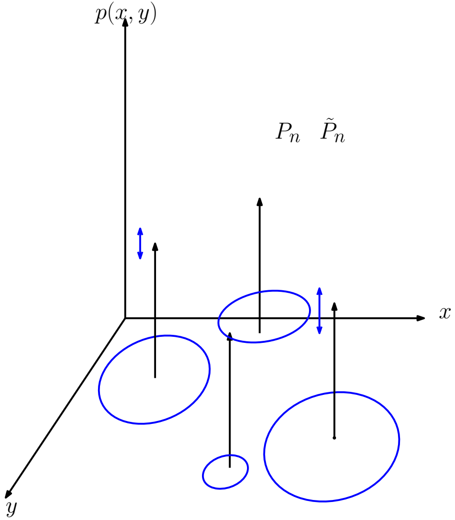
Robust Wasserstein profile function:

$$R_n(\beta) = \min\left\{ D_c(Q, P_n) : E_Q\left[\left(Y - \beta^T X\right)X\right] = 0 \right\}$$



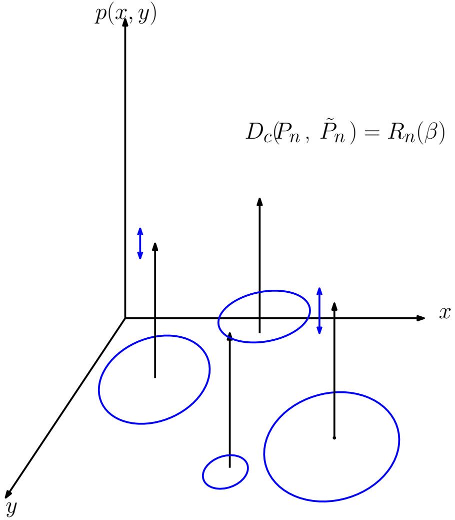
Robust Wasserstein profile function:

$$R_n(\beta) = \min\left\{D_c(Q, P_n) : E_Q\left[\left(Y - \beta^T X\right)X\right] = 0\right\}$$



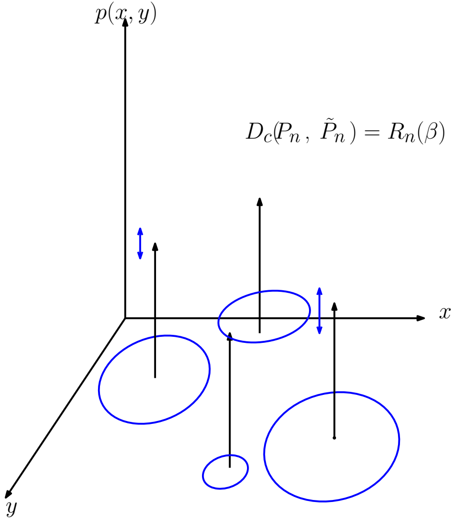
Robust Wasserstein profile function:

$$R_n(\beta) = \min\left\{D_c(Q, P_n) : E_Q\left[(Y - \beta^T X)X\right] = 0\right\}$$



Robust Wasserstein profile function:

$$R_n(\beta) = \min\left\{D_c\left(Q, P_n\right) : E_Q\left[\left(Y - \beta^T X\right)X\right] = 0\right\}$$



- Basically, $R_n(\beta)$ is a measure of goodness of β

$$nR_n(\beta) \longrightarrow \begin{cases} \mathcal{L}, & \text{if } \beta = \beta_* \\ \infty, & \text{if } \beta \neq \beta_* \end{cases}$$

- Similar to empirical likelihood profile function
- In high-dimensional setting, one can instead consider suitable non-asymptotic bounds for $nR_n(\beta)$.

RWPI Linear Regression:

$$\min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\leq\delta} E_Q\left[\left(Y-\beta^T X\right)^2\right]$$

←−−−−−− worst-case loss −−−−−−→

Theorem 2 [Blanchet, Kang & M]

If we take $c(u,v)=\|u-v\|_\infty^2$,

$$\text{Worst-case loss} = \left(\sqrt{\text{MSE}_n(\beta)}+\sqrt{\delta}\left\|\beta\right\|_1\right)^2$$

$$\text{Recall } D_c(P,Q)=\inf_\pi\left\{E_\pi\left[c(U,V)\right]:\pi_{_U}=P,\pi_{_V}=Q\right\}$$

RWPI Linear Regression:

$$\min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\leq\delta} E_Q\left[\left(Y-\beta^T X\right)^2\right]$$

$\longleftarrow$ worst-case loss $\longrightarrow$

Theorem 2 [Blanchet, Kang & M]

If we take $c(u,v)=\|u-v\|_\infty^2$,

$$\text{Worst-case loss} = \left(\sqrt{\text{MSE}_n(\beta)}+\sqrt{\delta}\|\beta\|_1\right)^2$$

$\Longrightarrow$ RWPI-Regression = Generalized Lasso!

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q : D_c(Q, P_n) \le \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

$\longleftarrow\!\!-\!-\!-\!-\!-$ worst-case loss $-\!-\!-\!-\!-\!\!\longrightarrow$

Theorem 2 [Blanchet, Kang & M]

If we take $c(u, v) = \|u - v\|_q^2$,

$$\text{Worst-case loss} = \left(\sqrt{\text{MSE}_n(\beta)} + \sqrt{\delta}\left\|\beta\right\|_p\right)^2$$

$\Longrightarrow$ RWPI-Regression(q) = ℓ_p-Penalized regression

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q\left[\left(Y - \beta^T X\right)^2\right]$$

$\longleftarrow\!\!-\!-\!-\!-\!-\!-$ worst-case loss $-\!-\!-\!-\!-\!-\!\!\longrightarrow$

Theorem 2 [Blanchet, Kang & M]

If we take $c(u, v) = \|u - v\|_q^2$,

$$\text{Worst-case loss} = \left(\sqrt{\text{MSE}_n(\beta)} + \sqrt{\delta}\left\|\beta\right\|_p\right)^2$$

A prescription for $\delta \Longrightarrow$ A prescription for regularization parameter

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \quad \underbrace{\max_{Q: D_c(Q, P_n) \leq \delta} \; E_Q\left|Y - \beta^T X\right|}_{\text{worst-case loss}}$$

Theorem 3 [Blanchet, Kang & M]

If we take $c(u, v) = \|u - v\|_q$,

$$\text{Worst-case loss} = \frac{1}{n} \sum_{i=1}^{n} \left|Y_i - \beta^T X_i\right| + \delta \|\beta\|_p$$

$\Longrightarrow$ RWPI linear regression with LAD loss = LAD - Lasso

RWPI Logistic Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q \left[\log \left(1 + exp(-Y \beta^T X) \right) \right]$$

←−−−−−− worst-case loss −−−−−−→

Theorem 3 [Blanchet, Kang & M]

If we take $c(u, v) = \|u - v\|_q^2$,

$$\text{Worst-case loss} = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + \exp(-Y_i \beta^T X_i) \right) + \delta \|\beta\|_p$$

$\Longrightarrow$ RWPI logistic regression = Penalized logistic regression

RWPI Hinge-loss minimization:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \le \delta} E_Q \left[\left(1 - Y \beta^T X\right)^+ \right]$$

$\longleftarrow$ worst-case loss $\longrightarrow$

Theorem 4 [Blanchet, Kang & M]

If we take $c(u, v) = \|u - v\|_q^2$,

$$\text{Worst-case loss} = \frac{1}{n} \sum_{i=1}^{n} \left(1 - Y_i \beta^T X_i\right)^+ + \delta \|\beta\|_p$$

$\Longrightarrow$ RWPI Hinge loss minimization $=$ SVM

Robust SAA:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D_c(Q, P_n) \leq \delta} E_Q \left[\text{Loss}(X, \beta) \right]$$

$\longleftarrow$———— worst-case loss ————$\longrightarrow$

Theorem 5 [Blanchet, Kang & M]

If we let $c(u, v) = \|u - v\|_2^2$ and $h(x, \beta) = D_\beta \text{Loss}(x, \beta)$,

$$R_n(\beta_*) \xrightarrow{D} \xi^T A^{-1} \xi,$$

where $\xi \sim \mathcal{N}(0, \text{Cov}[h(X, \beta_*)])$ and
$A = E\left[D_x h(X, \beta_*) D_x h(X, \beta_*)^T \right]$.

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \le \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

$$= \inf_{\beta \in \mathbb{R}^d} \left(\sqrt{MSE_n(\beta)} + \sqrt{\delta} \|\beta\|_1 \right)^2$$

A prescription for $\delta \implies$ A prescription for regularization parameter

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

$$= \inf_{\beta \in \mathbb{R}^d} \left(\sqrt{MSE_n(\beta)} + \sqrt{\delta} \|\beta\|_1 \right)^2$$

A prescription for $\delta \implies$ A prescription for regularization parameter

- Recall that we chose δ such that

$$P \left\{ R_n(\beta_*) \leq \delta \right\} \geq 1 - \alpha$$

- If X have sub-gaussian tails then, the corresponding prescription of tuning parameter turns out to be

$$c \frac{\Phi^{-1}(1 - \alpha/2d)}{\sqrt{n}} = O\left(\sqrt{\frac{\log d}{n}} \right)$$

Concluding remarks

- Distributional robustness
- Viewing regularization under the lens of distributional robustness
- Applications to stochastic optimization
- Additional learning applications where regularization structure may not be clear?....

RWPI Linear Regression:

$$\min_{\beta \in \mathbb{R}^d} \max_{Q: D(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

Model: $Y = 3X_1 + 2X_2 + 1.5X_4 + e$,
$X \sim \mathcal{N}(0, \Sigma)$, $\Sigma_{k,j} = 0.5^{|k-j|}$, $e \sim \mathcal{N}(0, 1)$
$n = 100$ training samples of (X, Y)

d	RWPI	Cross Validation	$(\log d/n)^{1/2}$
10	3 (3)	8 (3)	4 (3)
500	3 (3)	10 (3)	6 (3)
1000	3 (3)	19 (3)	11 (3)
3000	3 (3)	55 (3)	17 (3)

Table: Performance of different choices of regularization parameters for generalized Lasso.