Robust Wasserstein Profile Inference:

A new approach towards optimal regularization in machine learning

(Joint work with Jose Blanchet and Yang Kang)

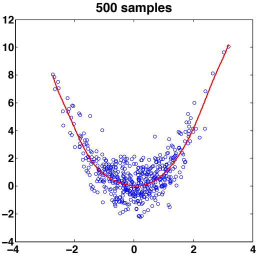
Karthyek Murthy
Columbia University

Bangalore Probability Seminar
April 2017

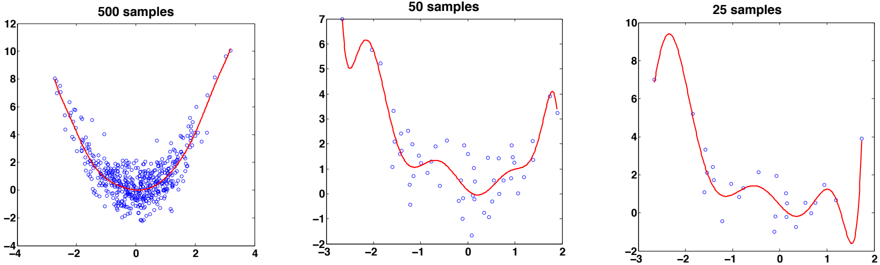
Objective

To use tools from robust stochastic optimization to avoid overfitting and systematically improve out of sample performance in statistical learning problems such as regression and classification.

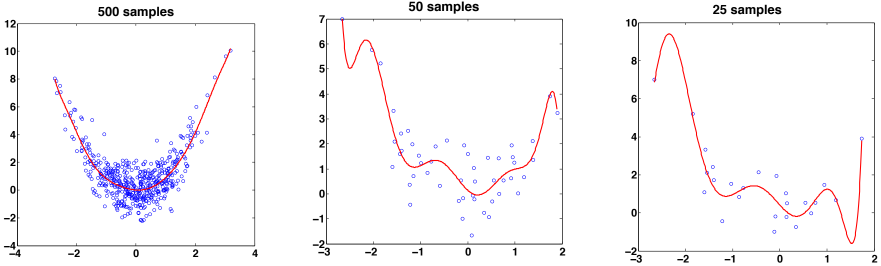
Overfitting - an illustration: n independent samples (X_i, Y_i) from the model $Y_i = X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim$ standard normal



Overfitting - an illustration: n independent samples (X_i, Y_i) from the model $Y_i = X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim$ standard normal



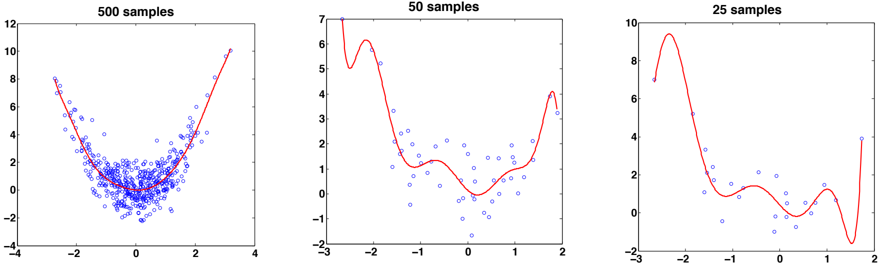
Overfitting - an illustration: n independent samples (X_i, Y_i) from the model $Y_i = X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim$ standard normal

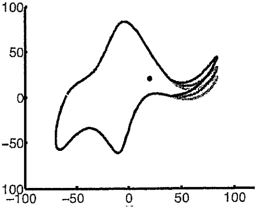


"With 4 parameters I can fit an elephant and with 5, I can make him wiggle his trunk."

- von Neumann

Overfitting - an illustration: n independent samples (X_i, Y_i) from the model $Y_i = X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim$ standard normal



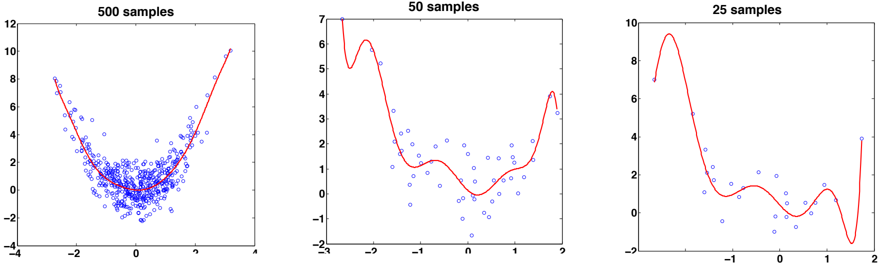


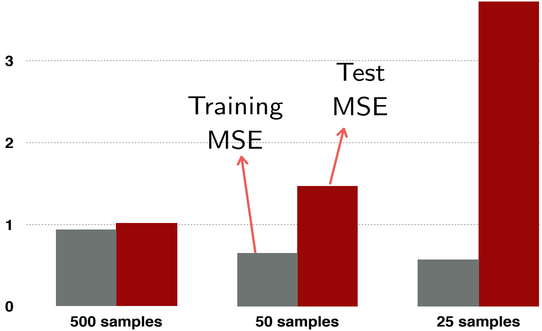
[Mayer et al '10]

"With 4 parameters I can fit an elephant and with 5, I can make him wiggle his trunk."

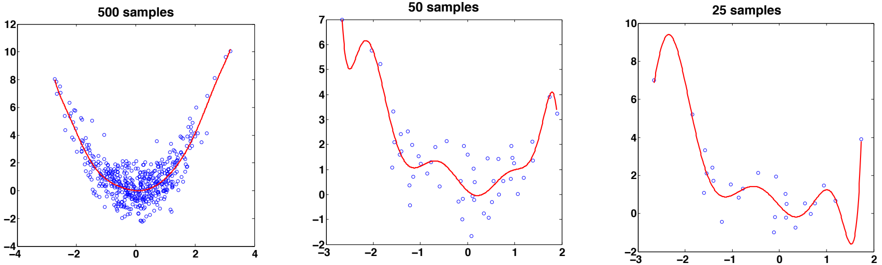
\- von Neumann

Overfitting - an illustration: n independent samples (X_i, Y_i) from the model $Y_i = X_i^2 + \varepsilon_i$, $\varepsilon_i \sim$ standard normal



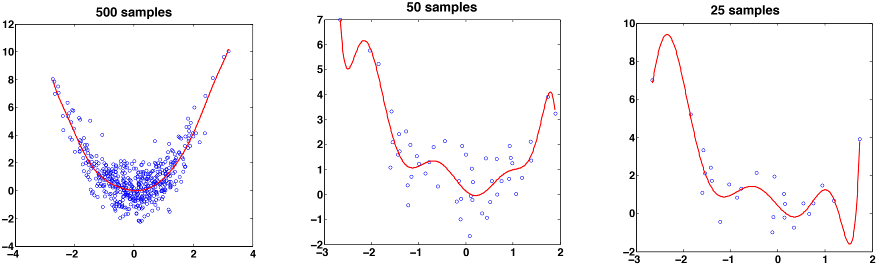


Overfitting - an illustration: $Y = X^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$



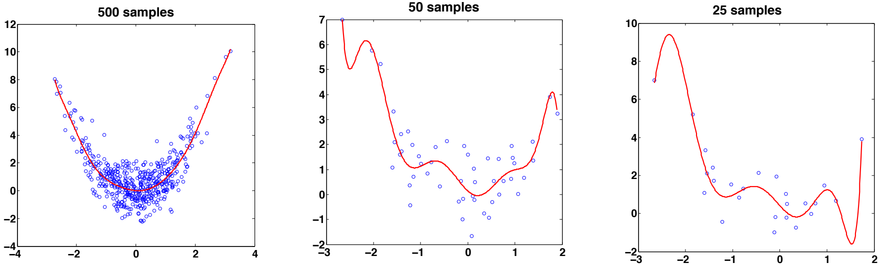
$$\min_{\beta} \quad \text{MSE}(\beta)$$

Overfitting - an illustration: $Y = X^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$



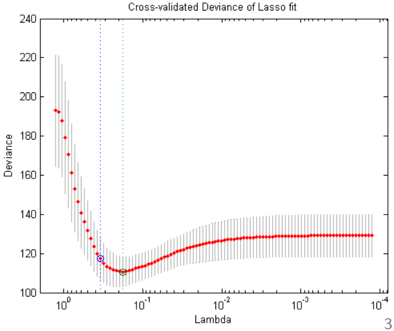
$$\min_{\beta} \left\{ \text{MSE}(\beta) + \underbrace{\lambda \|\beta\|}_{\text{regularization}} \right\}$$

Overfitting - an illustration: $Y = X^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$



$$\min_{\beta} \left\{ \text{MSE}(\beta) + \underbrace{\lambda \|\beta\|}_{\text{regularization}} \right\}$$

Cross-validation $\Longrightarrow$



travel adapter

All | Shopping | Images | Maps | News | More | Settings | Tools

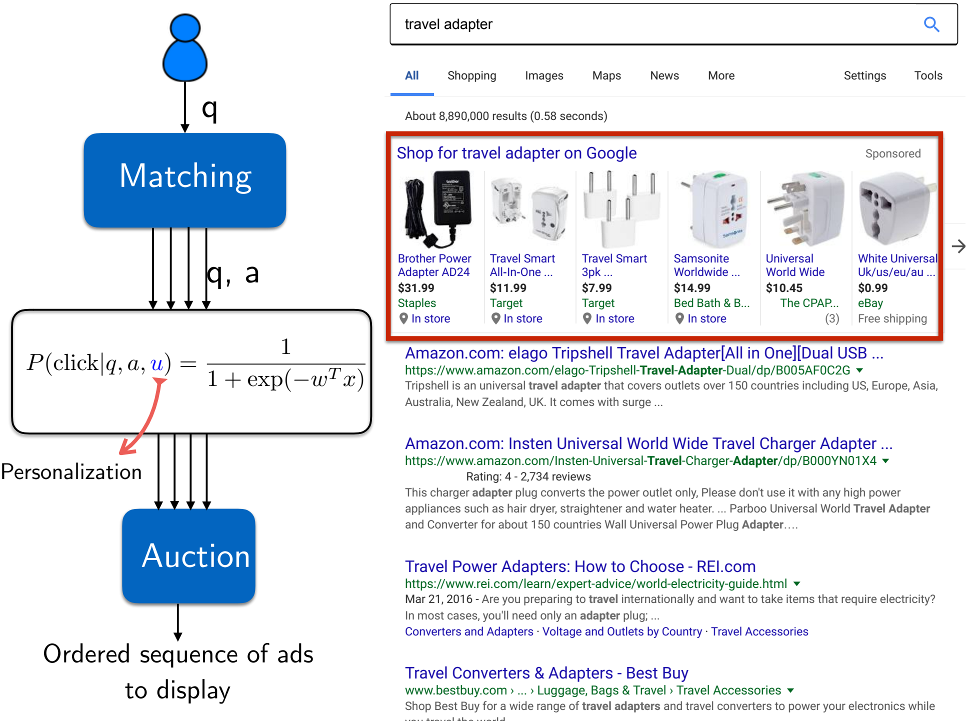
About 8,890,000 results (0.58 seconds)

Amazon.com: elago Tripshell Travel Adapter[All in One][Dual USB ...
https://www.amazon.com/elago-Tripshell-**Travel-Adapter**-Dual/dp/B005AF0C2G
Tripshell is an universal **travel adapter** that covers outlets over 150 countries including US, Europe, Asia, Australia, New Zealand, UK. It comes with surge ...

Amazon.com: Insten Universal World Wide Travel Charger Adapter ...
https://www.amazon.com/Insten-Universal-**Travel**-Charger-**Adapter**/dp/B000YN01X4
Rating: 4 - 2,734 reviews
This charger **adapter** plug converts the power outlet only, Please don't use it with any high power appliances such as hair dryer, straightener and water heater. ... Parboo Universal World **Travel Adapter** and Converter for about 150 countries Wall Universal Power Plug **Adapter**....

Travel Power Adapters: How to Choose - REI.com
https://www.rei.com/learn/expert-advice/world-electricity-guide.html
Mar 21, 2016 - Are you preparing to **travel** internationally and want to take items that require electricity? In most cases, you'll need only an **adapter** plug; ...
Converters and Adapters · Voltage and Outlets by Country · Travel Accessories

Travel Converters & Adapters - Best Buy
www.bestbuy.com › ... › Luggage, Bags & Travel › Travel Accessories
Shop Best Buy for a wide range of **travel adapters** and travel converters to power your electronics while you travel the world.

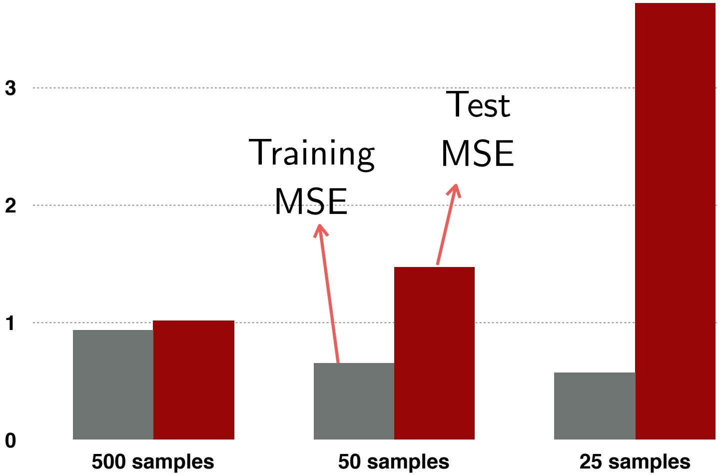

q
Matching
q, a
$P(\text{click}|q,a,u) = \frac{1}{1+\exp(-w^T x)}$
Personalization
Auction
Ordered sequence of ads to display
travel adapter
All
Shopping
Images
Maps
News
More
Settings
Tools
About 8,890,000 results (0.58 seconds)
Shop for travel adapter on Google
Sponsored
Brother Power Adapter AD24
$31.99
Staples
In store
Travel Smart All-In-One ...
$11.99
Target
In store
Travel Smart 3pk ...
$7.99
Target
In store
Samsonite Worldwide ...
$14.99
Bed Bath & B...
In store
Universal World Wide
$10.45
The CPAP...
(3)
White Universal Uk/us/eu/au ...
$0.99
eBay
Free shipping
Amazon.com: elago Tripshell Travel Adapter[All in One][Dual USB ...
https://www.amazon.com/elago-Tripshell-Travel-Adapter-Dual/dp/B005AF0C2G
Tripshell is an universal travel adapter that covers outlets over 150 countries including US, Europe, Asia, Australia, New Zealand, UK. It comes with surge ...
Amazon.com: Insten Universal World Wide Travel Charger Adapter ...
https://www.amazon.com/Insten-Universal-Travel-Charger-Adapter/dp/B000YN01X4
Rating: 4 - 2,734 reviews
This charger adapter plug converts the power outlet only, Please don't use it with any high power appliances such as hair dryer, straightener and water heater. ... Parboo Universal World Travel Adapter and Converter for about 150 countries Wall Universal Power Plug Adapter....
Travel Power Adapters: How to Choose - REI.com
https://www.rei.com/learn/expert-advice/world-electricity-guide.html
Mar 21, 2016 - Are you preparing to travel internationally and want to take items that require electricity? In most cases, you'll need only an adapter plug; ...
Converters and Adapters · Voltage and Outlets by Country · Travel Accessories
Travel Converters & Adapters - Best Buy
www.bestbuy.com › ... › Luggage, Bags & Travel › Travel Accessories
Shop Best Buy for a wide range of travel adapters and travel converters to power your electronics while

$\sqrt{\texttt{Lasso}}$

$$\min_{\beta} \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta^T \mathbf{x_i}\right)^2} \quad + \quad \lambda\|\beta\|_1$$

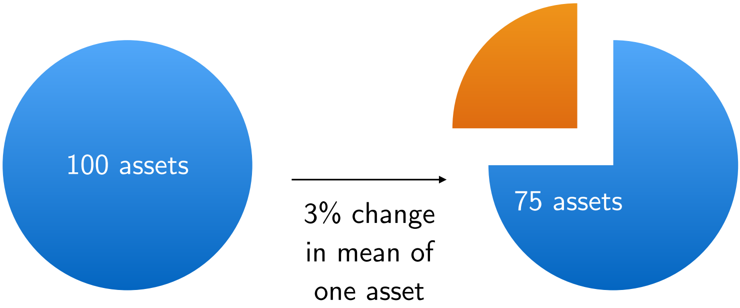
Regularized logisitc regression

$$\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp(-y_i\beta^T \mathbf{x_i})\right) \quad + \quad \lambda\|\beta\|_1$$


3
2
1
0
Training
MSE
Test
MSE
500 samples
50 samples
25 samples

"The optimizer's curse"

“The optimizer’s curse”



[Best & Grauer ’91]

The premise of distributionally robust optimization

To solve:

$$\min_{\beta} \; E\left[\mathrm{Loss}(W; \beta)\right]$$

ERM / SAA:

$$\min_{\beta} \; \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}(W_i; \beta)$$

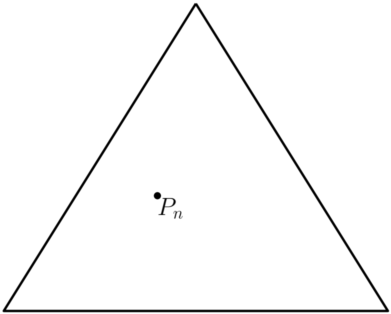
The premise of distributionally robust optimization

To solve:

$$\min_{\beta} \; E\left[\mathrm{Loss}(W;\beta)\right]$$

ERM / SAA:

$$\min_{\beta} \; \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}(W_i;\beta)$$



The premise of distributionally robust optimization

To solve:

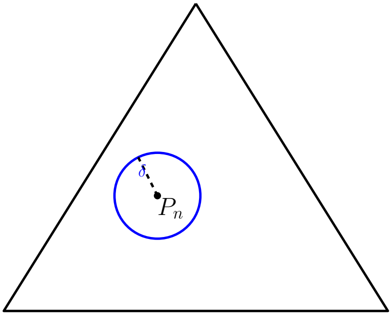
$$\min_{\beta} \; E\left[\text{Loss}(W;\beta)\right]$$

ERM / SAA:

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

DRO:

$$\min_{\beta} \; \max_{Q:D(Q,P_n)\leq\delta} E_Q\left[\text{Loss}(W;\beta)\right]$$



The premise of distributionally robust optimization

To solve:

$$\min_{\beta} \; E\left[\text{Loss}(W;\beta)\right]$$

ERM / SAA:

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

DRO:

$$\min_{\beta} \; \max_{Q:D(Q,P_n)\leq\delta} E_Q\left[\text{Loss}(W;\beta)\right]$$

Example 1

DR linear regression:

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

The premise of distributionally robust optimization

To solve:

$$\min_{\beta} \; E\left[\text{Loss}(W;\beta)\right]$$

ERM / SAA:

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

DRO:

$$\min_{\beta} \; \max_{Q: D(Q,P_n)\le\delta} E_Q\left[\text{Loss}(W;\beta)\right]$$

Example 1

DR linear regression:

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

Objective

- Improve generalization with DRO
- self-tune?

The premise of distributionally robust optimization

Example 1

DR linear regression:

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

Objective

- Improve generalization with DRO
- self-tune?

Q1) How to quantify D?

Q2) How to choose δ?

Outline of rest of the presentation

- Motivation
- The distributionally robust approach
  - The premise of DRO
- Q1) How to choose the distance function
  - Optimal transport based DRO formulation
- Q2) How to choose the tuning parameter?
  - Profile function
  - Tuning parameter as a quantile of the profile function
- Discussion

DR Linear Regression:

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

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

DR Linear Regression:

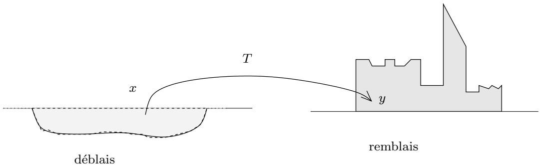
$$\min_{\beta \in \mathbb{R}^d} \max_{Q : D(Q, P_n) \leq \delta} E_Q \left[\left(Y - \beta^T X \right)^2 \right]$$

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

$$D(P, Q) = \min_{\pi : \pi_U = P, \pi_V = Q} E_\pi \| U - V \|$$

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]

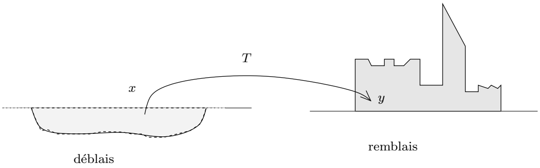
How to quantify the distance $D(P, Q)$?

$$D(P, Q) = \min_{\pi : \pi_U = P, \pi_V = Q} E_\pi \| U - V \|$$

[1] Image source: Optimal Transport: Old and New by Cédric Villani

DR Linear Regression:

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



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

The metric D_c is called optimal transport metric.
When $c(u, v) = \|u - v\|^\rho$, $D_c^{1/\rho}$ is the ρ^{th} order Wasserstein distance

Why optimal transport distances?

$$\mathcal{P} = \{P : D_{KL}(P \| P_{ref}) \leq \delta\}$$

Hansen and Sargent '01, '06

Nilim and El Ghaoui '02, '03

Iyengar '05

Lim, Shanthikumar and Watewai '05, '06

Jain, Lim and Shanthikumar '10

Ben-Tal et al '13

Lam '13, '16

Csiszàr and Breuer '13

Jiang and Guan '12

Hu and Hong '13

Wang, Glynn and Ye '14

Glasserman and Xu '14

Bayrakskan and Love '15

Shapiro '15

Duchi, Glynn and Namkoong '16

Dhara, Das and Natarajan '17

Why optimal transport distances?

$$\mathcal{P} = \left\{P : D_{KL}(P \| P_{ref}) \leq \delta\right\}$$

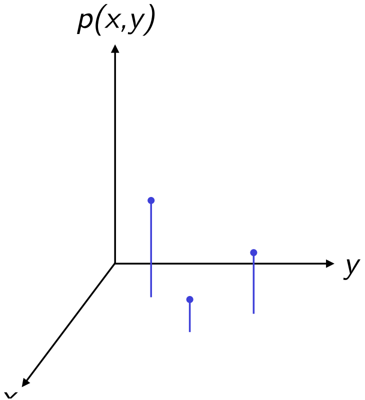
$$D_{KL}(p\|q) = \begin{cases} \int p(x) \log \frac{p(x)}{q(x)} dx & \text{if } p \ll q \\ \infty & \text{otherwise.} \end{cases}$$

Why optimal transport distances?

$$\mathcal{P} = \left\{ P : D_{KL}(P \| P_{ref}) \leq \delta \right\}$$

$$D_{KL}(p\|q) = \begin{cases} \int p(x) \log \frac{p(x)}{q(x)} dx & \text{if } p \ll q \\ \infty & \text{otherwise.} \end{cases}$$

Baseline probability distribution p

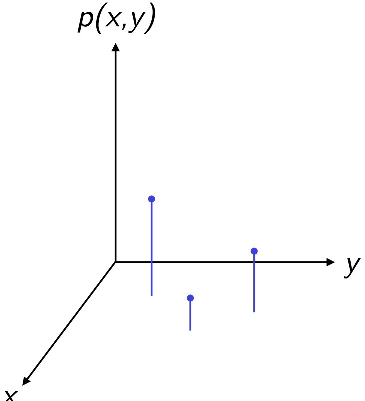


Why optimal transport distances?

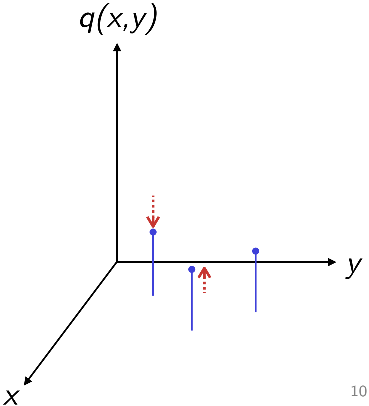
$$\mathcal{P} = \{P : D_{KL}(P \| P_{ref}) \leq \delta\}$$

$$D_{KL}(p\|q) = \begin{cases} \int p(x) \log \frac{p(x)}{q(x)} dx & \text{if } p \ll q \\ \infty & \text{otherwise.} \end{cases}$$

Baseline probability distribution p



A KL-neighbor of p

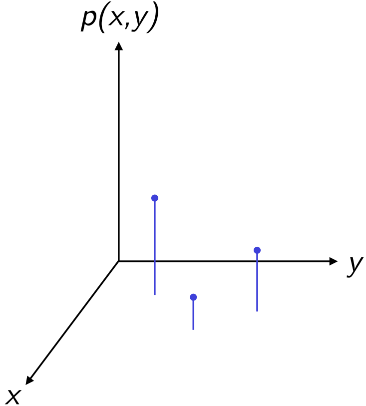


Why optimal transport distances?

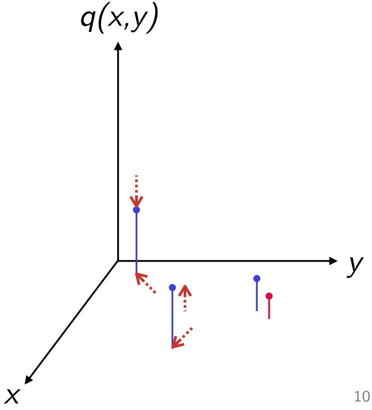
$$\mathcal{P} = \{P : D_{KL}(P\|P_{ref}) \leq \delta\}$$

$$D_{KL}(p\|q) = \begin{cases} \int p(x)\log\frac{p(x)}{q(x)}dx & \text{if } p \ll q \\ \infty & \text{otherwise.} \end{cases}$$

Baseline probability distribution p



A Wasserstein neighbor of p



Application 1: Linear regression

OLS:

$$\min_{\beta \in \mathbb{R}^d} \ \mathrm{MSE}_n(\beta) \quad \overset{\text{DRO}}{- - \longrightarrow}$$

DR linear regression:

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

Application 1: Linear regression

OLS:

$$\min_{\beta\in\mathbb{R}^d} \mathrm{MSE}_n(\beta)$$

$$\xrightarrow{\text{DRO}}$$

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

$$\arg\min_{\beta} \sup_{Q:D_c(Q,P_n)\le\delta} E_P\left[(Y-\beta^T X)^2\right]$$

$$=$$

Application 1: Linear regression

OLS:

$$\min_{\beta \in \mathbb{R}^d} \ \mathrm{MSE}_n(\beta)$$

$$\xrightarrow{\text{DRO}}$$

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

$$\arg\min_{\beta} \ \sup_{Q : D_c(Q, P_n) \le \delta} \ E_P\left[(Y - \beta^T X)^2\right] = \arg\min_{\beta} \left\{ \sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta} \|\beta\|_p \right\}$$

DR-linear regression $= \ell_p$-penalized regression!

Application 1: Linear regression

OLS:

$$\min_{\beta\in\mathbb{R}^d} \mathrm{MSE}_n(\beta) \quad \xrightarrow{\text{DRO}} $$

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: If $c(u,v)=\|u-v\|_q^2$,
(Recall $D_c(P,Q)=\min E[c(U,V)]$)

$$\arg\min_{\beta} \sup_{Q:D_c(Q,P_n)\le\delta} E_P\left[(Y-\beta^T X)^2\right] = \arg\min_{\beta}\left\{\sqrt{\mathrm{MSE}_n(\beta)}+\sqrt{\delta}\|\beta\|_p\right\}$$

DR-linear regression $=$ ℓ_p-penalized regression!

Application 1: Linear regression

`OLS:`

$$\min_{\beta \in \mathbb{R}^d} \mathrm{MSE}_n(\beta) \quad \xrightarrow{\text{DRO}} \quad$$

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

$$\arg\min_{\beta} \sup_{Q : D_c(Q, P_n) \le \delta} E_P\left[(Y - \beta^T X)^2\right] = \arg\min_{\beta} \left\{ \sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta} \|\beta\|_1 \right\}$$

DR-linear regression $= \sqrt{\text{Lasso}}$!

Application 2: Logistic regression

ERM:

DR linear regression:

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \text{Logistic loss}(X_i;\beta) \xrightarrow{\text{DRO}} \min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\le\delta} E_Q\left[\text{Logistic loss}(X;\beta)\right]$$

Theorem: If $c(u,v) = \|u-v\|_q$,

$$\arg\min_{\beta} \sup_{Q:D_c(Q,P_n)\le\delta} E_P\left[\text{Logistic loss}(X;\beta)\right]$$

$$= \arg\min_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\text{Logistic loss}(X_i;\beta) + \delta\|\beta\|_p\right\}$$

DR-logistic regression $=$ ℓ_p-penalized logistic regression!

Pflug et al (2012)

Wozabal (2012)

Lee and Mehrotra (2013),

Kuhn et al (2015)

Blanchet & M (2016)

Gao & Kleywegt (2016)

Duality theorem [Blanchet & M]

S Polish space

$P_{ref} \in P(S)$ reference measure

$f \in L^1(dP_{ref})$ is upper semicontinuous

$\delta \in (0, \infty)$

A lower semicontinuous cost function $c : S \times S \to \mathbb{R}$ satisfying $c(x, x) = 0$ for all $x \in S$.

Duality holds

$$\sup\left\{ \int f dP : d_c(P, P_{ref}) \leq \delta \right\} = \inf_{\lambda \geq 0} \left\{ \lambda\delta + E_{ref}\left[\sup_{y \in S} \left\{ f(y) - \lambda c(X, y) \right\} \right] \right\}$$

Outline

- Motivation
- The distributionally robust approach

  The premise of DRO
- Q1) How to choose the distance function ✓

  Optimal transport based DRO formulation
- Q2) How to choose the tuning parameter?

  Profile function

  Tuning parameter as a quantile of the profile function
- Discussion

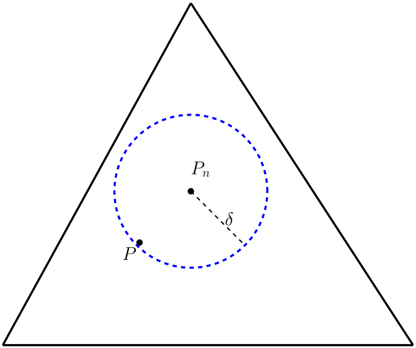
DR Linear Regression:

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

How do we choose δ?

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 do we choose δ?

$$P\left(D_c(P, P_n) \leq \delta\right) \geq 1 - \varepsilon$$

See Fournier and Guillin (2015)

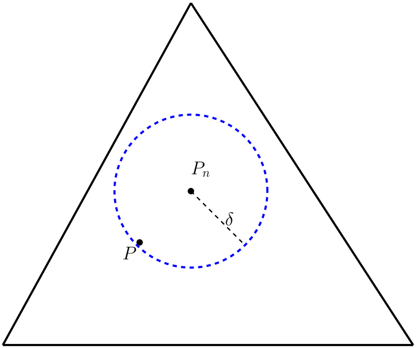
Lee and Mehrotra (2013), Kuhn et al (2015), $O(n^{-1/d})$ rate

DR Linear Regression:

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

Given Q,
$\beta_{(Q)} :=$ optimal β satisfying

$$E_Q \left[\left(Y - \beta_{(Q)}^T X \right) X \right] = \mathbf{0}$$



DR Linear Regression:

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

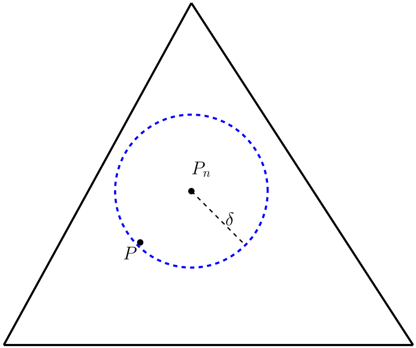
Plausible β's:

$$\beta_* \in \left\{\beta_{(Q)} : D_c(Q, P_n) \le \delta\right\}$$

Given Q,

$\beta_{(Q)} :=$ optimal β satisfying

$$E_Q\left[\left(Y - \beta_{(Q)}^T X\right) X\right] = \mathbf{0}$$



DR Linear Regression:

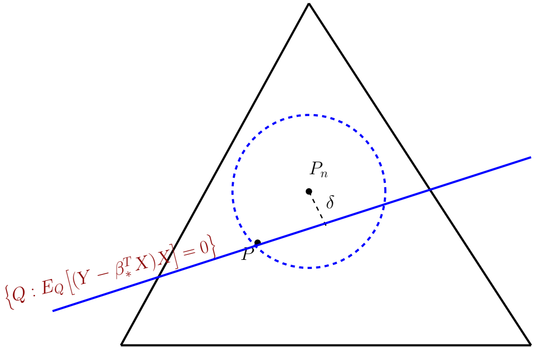
$$\min_{\beta\in\mathbb{R}^d} \max_{Q:D_c(Q,P_n)\leq\delta} E_Q\left[\left(Y-\beta^T X\right)^2\right]$$

Plausible β's:

$\beta_* \in \left\{\beta_{(Q)} : D_c(Q,P_n) \leq \delta\right\}$

β_* is the optimal β satisfying

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

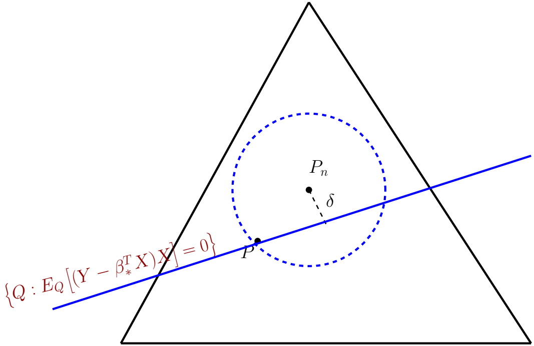


DR Linear Regression:

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

Plausible β's:

$\beta_* \in \left\{\beta_{(Q)} : D_c(Q, P_n) \leq \delta\right\}$



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

DR Linear Regression:

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

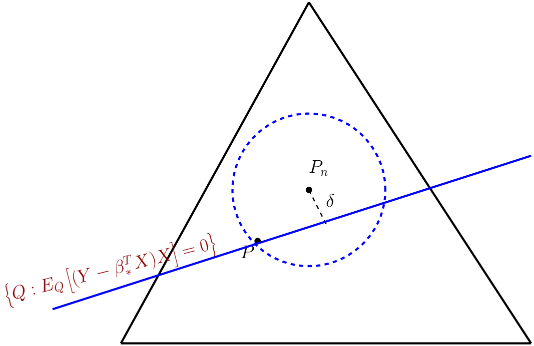
Plausible β's:

$$\beta_* \in \left\{\beta_{(Q)} : D_c(Q, P_n) \leq \delta\right\}$$

Theorem

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

$$nR_n(\beta_*) \xrightarrow{D} \bar{R}$$



Choose $\delta = \frac{\eta}{n}$ where η is such that $P\left\{\bar{R} \leq \eta\right\} \geq 0.95$

DR Linear Regression:

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

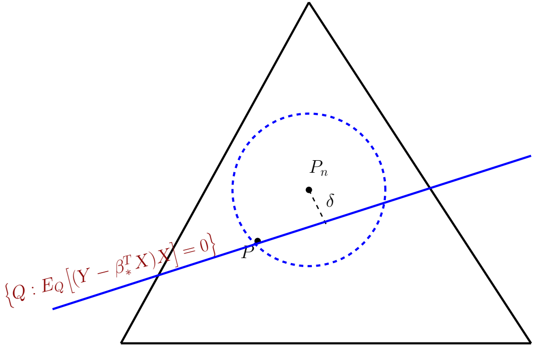
Plausible β's:

$$\beta_* \in \left\{\beta_{(Q)} : D_c(Q, P_n) \leq \delta\right\}$$

Theorem

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

$$nR_n(\beta_*) \xrightarrow{D} \bar{R}$$



Choose $\delta = \frac{\eta_\alpha}{n}$ where η_α is such that $P\left\{\bar{R} \leq \eta_\alpha\right\} = 1 - \alpha$.

DR Linear Regression:

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

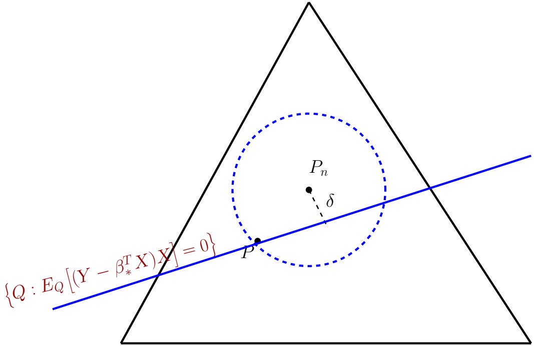
Plausible β's:

$$\beta_* \in \left\{\beta_{(Q)} : D_c(Q, P_n) \le \delta\right\}$$

Theorem

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

$$nR_n(\beta_*) \xrightarrow{D} \bar{R}$$



Then $P\left(\beta_* \in \text{Plausible set}\right) \approx 1 - \alpha$.

`Optimality condition:` $E\left[h(W;\beta_*)\right] = \mathbf{0}$

`RWP function:` $R_n(\beta) = \inf\left\{D_c(Q, P_n) : E_Q\left[h(W,\beta)\right] = \mathbf{0}\right\}$

- Similar to empirical likelihood profile function

$$T(\beta) = \max\left\{\sum_{i=1}^{n} \log p_i : \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i h(w_i, \beta) = \mathbf{0}\right\}$$

Optimality condition: $E\left[h(W;\beta_*)\right] = \mathbf{0}$

RWP function: $R_n(\beta) = \inf\left\{D_c(Q, P_n) : E_Q\left[h(W,\beta)\right] = \mathbf{0}\right\}$

- Similar to empirical likelihood profile function

$$T(\beta) = \max\left\{\sum_{i=1}^{n} \log\frac{p_i}{1/n} : \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i h(w_i, \beta) = \mathbf{0}\right\}$$

$$= \min\left\{D_{\mathrm{KL}}(Q \| P_n) : E_Q[h(W_i, \theta_*)] = \mathbf{0}\right\}$$

- $T(\beta_*)$ typically has a χ^2-limiting distribution

Optimality condition: $E\left[h(W;\beta_*)\right] = \mathbf{0}$

RWP function: $R_n(\beta) = \inf\left\{D_c(Q, P_n) : E_Q\left[h(W,\beta)\right] = \mathbf{0}\right\}$

Theorem

If we let $c(u,v) = \|u - v\|_q^\rho$,

$$n^{\rho/2} R_n\left(\beta_*\right) \xrightarrow{D} \bar{R},$$

$$\bar{R} = \sup_{\zeta \in \mathbb{R}^r} \left\{\rho\zeta^T Z - (\rho - 1) E\left\|\zeta^T D_w h\left(W, \beta_*\right)\right\|_p^{\rho/(\rho-1)}\right\}$$

Optimality condition: $E\left[h(W;\beta_*)\right] = \mathbf{0}$

RWP function: $R_n(\beta) = \inf\left\{D_c(Q,P_n) : E_Q\left[h(W,\beta)\right] = \mathbf{0}\right\}$

Theorem

If we let $c(u,v) = \|u-v\|_q^\rho$,

$$\boxed{n^{\rho/2} R_n\left(\beta_*\right) \xrightarrow{D} \bar{R},}$$

ℓ_p-lin reg: $\rho = 2$

$$\bar{R} \overset{D}{\leq} \frac{\pi}{\pi-2}\|Z\|_q^2,$$

ℓ_p-log reg: $\rho = 1$

$$\bar{R} \overset{D}{\leq} \|Z\|_q,$$

where $Z \sim \mathcal{N}(\mathbf{0}, E[XX^T])$.

$$\bar{R} = \sup_{\zeta\in\mathbb{R}^r}\left\{\rho\zeta^T Z - (\rho-1)E\left\|\zeta^T D_w h\left(W,\beta_*\right)\right\|_p^{\rho/(\rho-1)}\right\}$$

Optimality condition: $E\left[h(W;\beta_*)\right] = \mathbf{0}$

RWP function: $R_n(\beta) = \inf\left\{D_c(Q, P_n) : E_Q\left[h(W,\beta)\right] = \mathbf{0}\right\}$

Theorem

If we let $c(u,v) = \|u - v\|_q^\rho$,

$$\boxed{n^{\rho/2} R_n\left(\beta_*\right) \xrightarrow{D} \bar{R},}$$

ℓ_p−lin reg: $\rho = 2$

$$\bar{R} \overset{D}{\leq} \frac{\pi}{\pi - 2}\|Z\|_q^2,$$

ℓ_p−log reg: $\rho = 1$

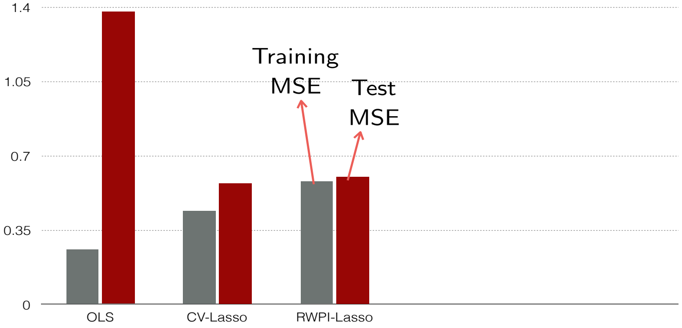
$$\bar{R} \overset{D}{\leq} \|Z\|_q,$$

where $Z \sim \mathcal{N}(\mathbf{0}, E[XX^T])$.

$$nR_n(\beta_*) \leq \frac{\pi}{\pi - 2}\frac{\Phi^{-1}\left(1 - \alpha/2d\right)}{\sqrt{n}} = O\left(\sqrt{\frac{\log d}{n}}\right)$$

RWPI Linear Regression:

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



RWPI based tuning parameter selection against cross-validated Lasso and OLS in the diabetes data set of 142 training samples with 64 predictors

A snapshot of main results

Application 1: DR linear regression

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

$$\arg\min_{\beta} \sup_{Q: D_c(Q,P_n) \le \delta} E_P\left[(Y - \beta^T X)^2\right]$$

$$= \arg\min_{\beta} \left\{ \sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta}\|\beta\|_p \right\}$$

$$\sqrt{\frac{\pi}{\pi - 2}} \frac{\|Z\|_q}{\sqrt{n}}$$

Application 2: DR logistic regression

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

$$\arg\min_{\beta} \sup_{Q: D_c(Q,P_n) \le \delta} E_P[\text{Logistic loss}(X;\beta)]$$

$$= \arg\min_{\beta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \text{Logistic loss}(X_i;\beta) + \delta\|\beta\|_p \right\}$$

$$\frac{\|Z\|_q}{\sqrt{n}}$$

A snapshot of main results

Application 1: `DR linear regression`

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

$$\arg\min_{\beta} \sup_{Q: D_c(Q,P_n) \le \delta} E_P\left[(Y - \beta^T X)^2\right]$$

$$= \arg\min_{\beta} \left\{ \sqrt{\mathrm{MSE}_n(\beta)} + \sqrt{\delta} \|\beta\|_p \right\}$$

$$\sqrt{\frac{\pi}{\pi - 2}} \frac{\|Z\|_q}{\sqrt{n}}$$

Application 2: `DR logistic regression`

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

$$\arg\min_{\beta} \sup_{Q: D_c(Q,P_n) \le \delta} E_P\left[\text{Logistic loss}(X;\beta)\right]$$

$$= \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{Logistic loss}(X_i;\beta) + \delta \|\beta\|_p \right\}$$

$$\frac{\|Z\|_q}{\sqrt{n}}$$

- Scalability: Similar equivalences with SVM, LAD-regression

Discussion

- DRO approach towards improving out-of-sample performance

Discussion

- DRO approach towards improving out-of-sample performance
- Optimal uncertainty size as a notion of plausibility

Discussion

- DRO approach towards improving out-of-sample performance
- Optimal uncertainty size as a notion of plausibility
- Popular regularized estimators as particular cases

Discussion

- DRO approach towards improving out-of-sample performance
- Optimal uncertainty size as a notion of plausibility
- Popular regularized estimators as particular cases
- A partial answer to "why optimal transport based distances?"

Discussion

- DRO approach towards improving out-of-sample performance
- Optimal uncertainty size as a notion of plausibility
- Popular regularized estimators as particular cases
- A partial answer to "why optimal transport based distances?"
- Potential to generate new algorithms that self-tune and systematically improve out-of-sample-performance

Discussion

- DRO approach towards improving out-of-sample performance
- Optimal uncertainty size as a notion of plausibility
- Popular regularized estimators as particular cases
- A partial answer to “why optimal transport based distances?”
- Potential to generate new algorithms that self-tune and systematically improve out-of-sample-performance
- Future research: Optimal choice of cost functions, computational methods, multivariate extremes, etc.

Paper: *Robust Wasserstein Profile Inference* (Available in arXiv)