DRO with optimal transport distances:
Some statistical and algorithmic advances

(joint work with Jose Blanchet, Yang Kang & Fan Zhang)

Karthyek Murthy

Singapore University of Technology and Design

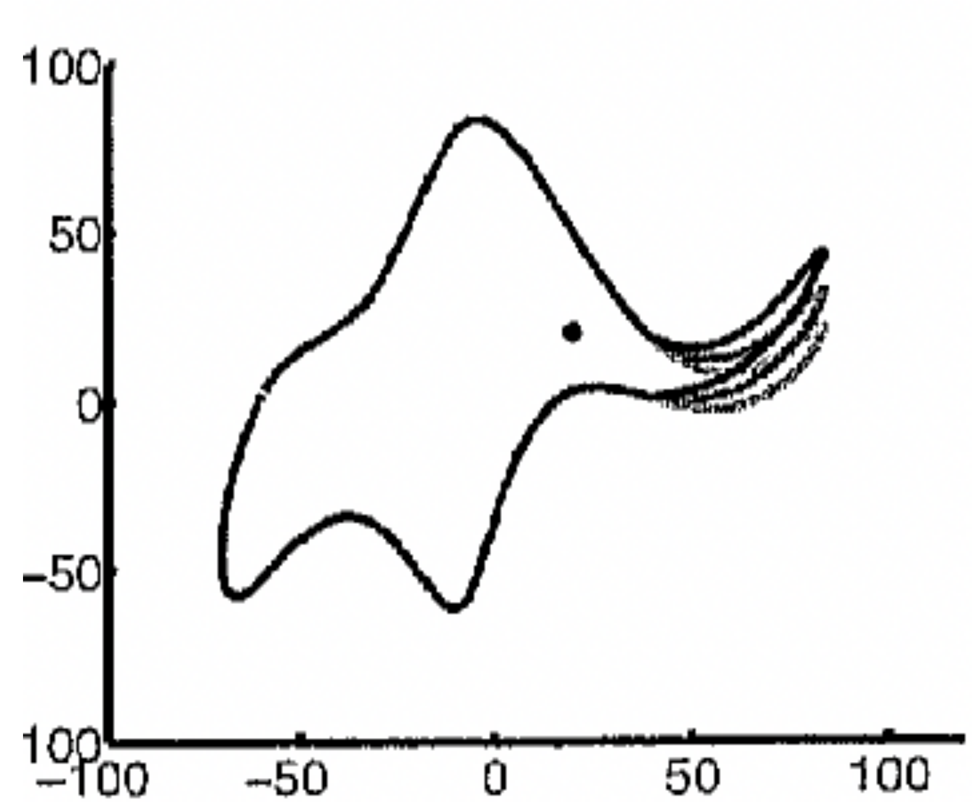
DRO meet, Banff

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

-von Neumann

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

-von Neumann



Mayer et al '10

$$\inf_{\beta} \quad \sup_{P \in \mathcal{P}} \quad E_P \left[\ell(X; \beta) \right]$$

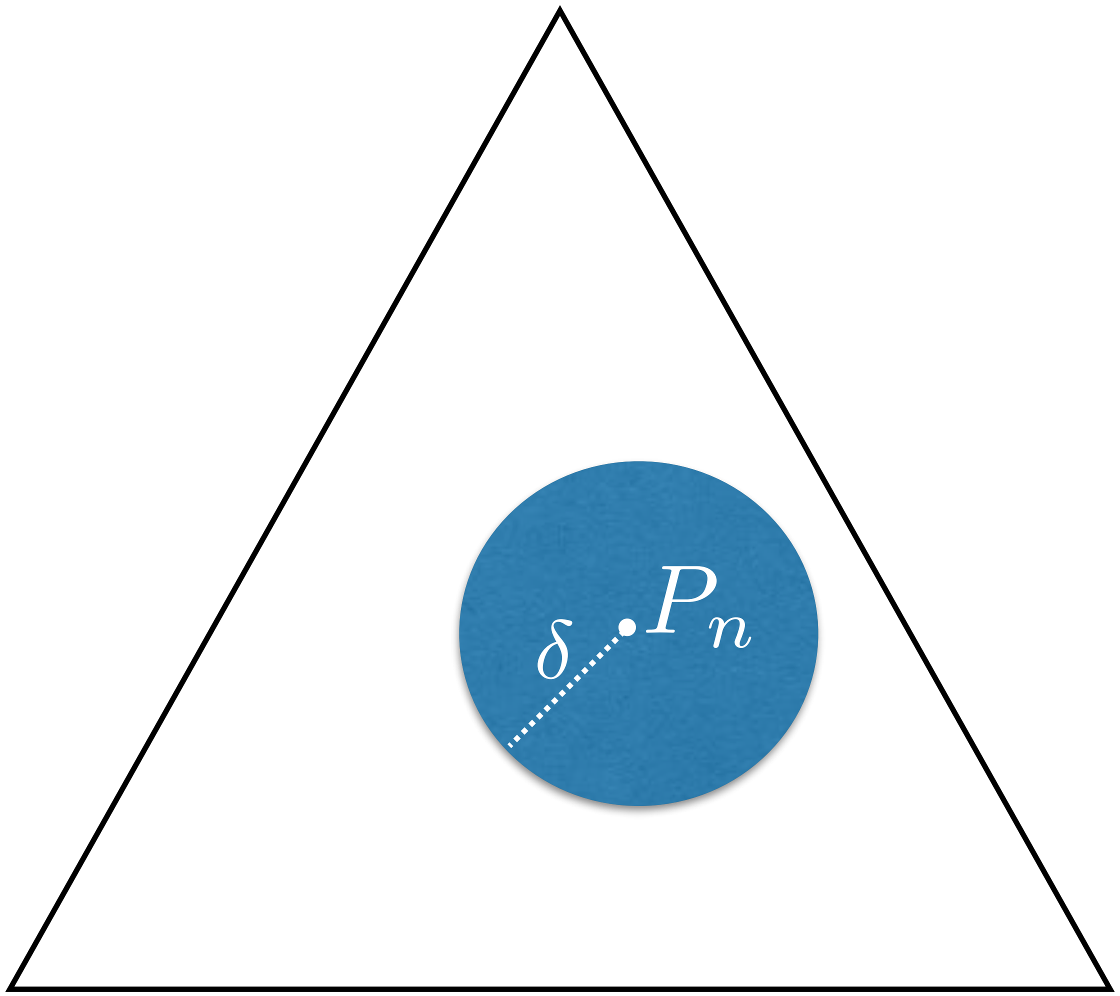
Specifying the set of plausible distributions $\mathcal{P}$:

- Moment assumptions
- Structural assumptions (unimodal, convex tails,...)
- Statistical/probabilistic distances

$$\inf_{\beta} \sup_{P:D(P,P_n)\leq\delta} E_P\left[\ell(X;\beta)\right]$$

Specifying the set of plausible distributions $\mathcal{P}$:

- Moment assumptions
- Structural assumptions (unimodal, convex tails,...)
- Statistical/probabilistic distances



$$\inf_{\beta} \sup_{P:D(P,P_n)\leq\delta} E_P\left[\ell(X;\beta)\right]$$

Specifying the set of plausible distributions $\mathcal{P}$:

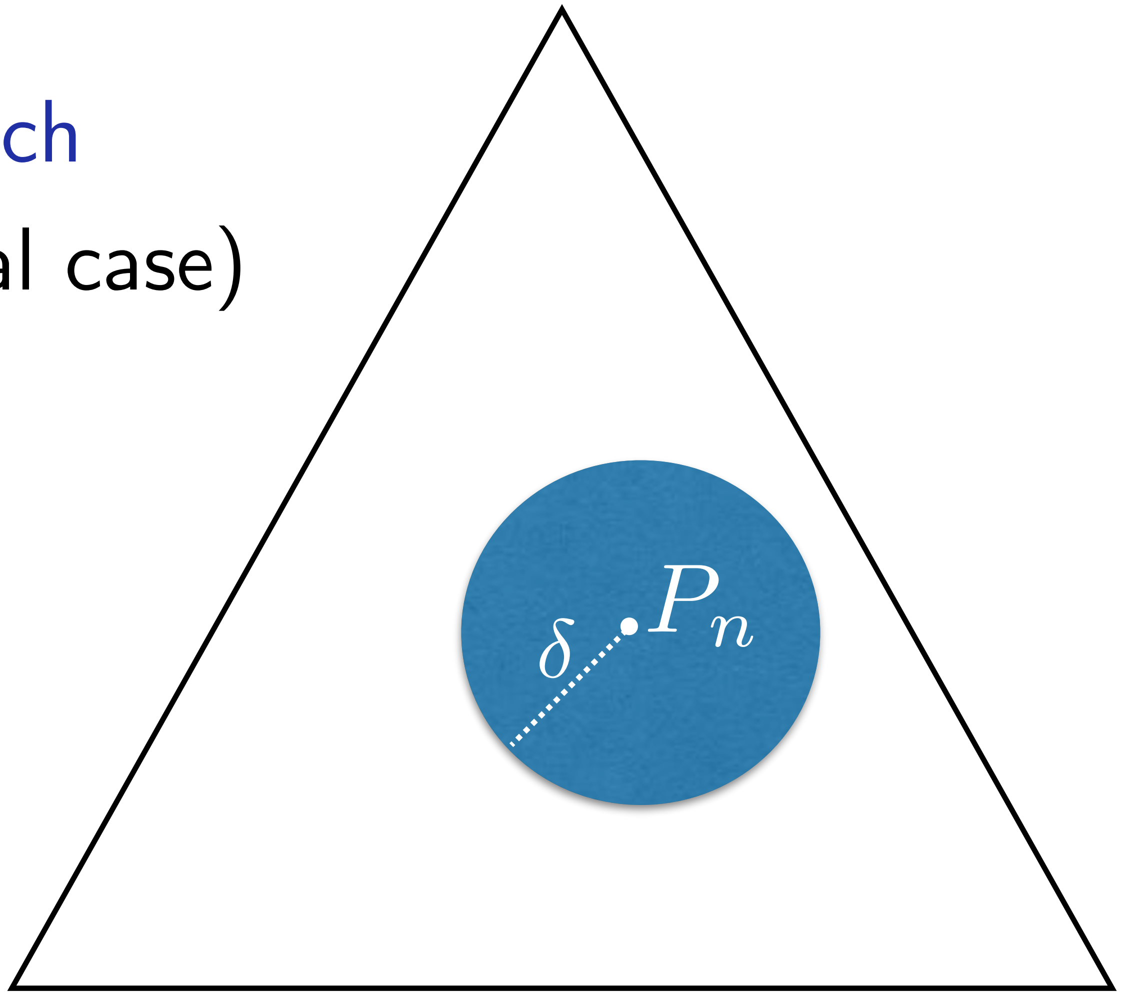
Moment assumptions

Structural assumptions (unimodal, convex tails,...)

Statistical/probabilistic distances

→ optimal transport based approach

(includes Wasserstein DRO as a special case)



$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a powerful & flexible tool towards introducing model ambiguity in data-driven optimization under uncertainty

$$\inf_{\beta} \sup_{P: D(P,P_n) \leq \delta} E_P[\ell(X;\beta)] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a powerful & flexible tool towards introducing model ambiguity in data-driven optimization under uncertainty

→ A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO)

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a powerful & flexible tool towards introducing model ambiguity in data-driven optimization under uncertainty

→ A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO)

→ Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

$$\inf_{\beta} \sup_{P: D(P, P_n) \le \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a powerful & flexible tool towards introducing model ambiguity in data-driven optimization under uncertainty

→ A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO)

→ Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

A Stochastic gradient descent scheme that is at least "as fast", or sometimes much faster than the non-robust counterpart!

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P[\ell(X;\beta)] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

- → A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO)
- → Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?
- → How do we specify the parameters for the ambiguity model?

$$\inf_{\beta} \sup_{P:D(P,P_n)\leq\delta} E_P\left[\ell(X;\beta)\right] \qquad \text{(OT-DRO)}$$

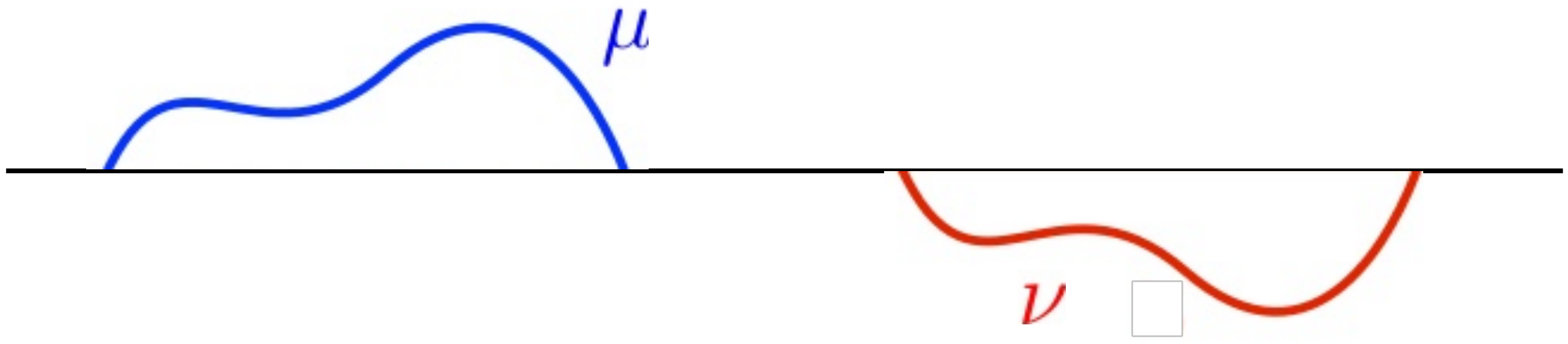
Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

- A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO)
- Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?
- How do we specify the parameters for the ambiguity model?
  - choosing the radius
  - utilising data to inform the geometry of the ambiguous neighborhood

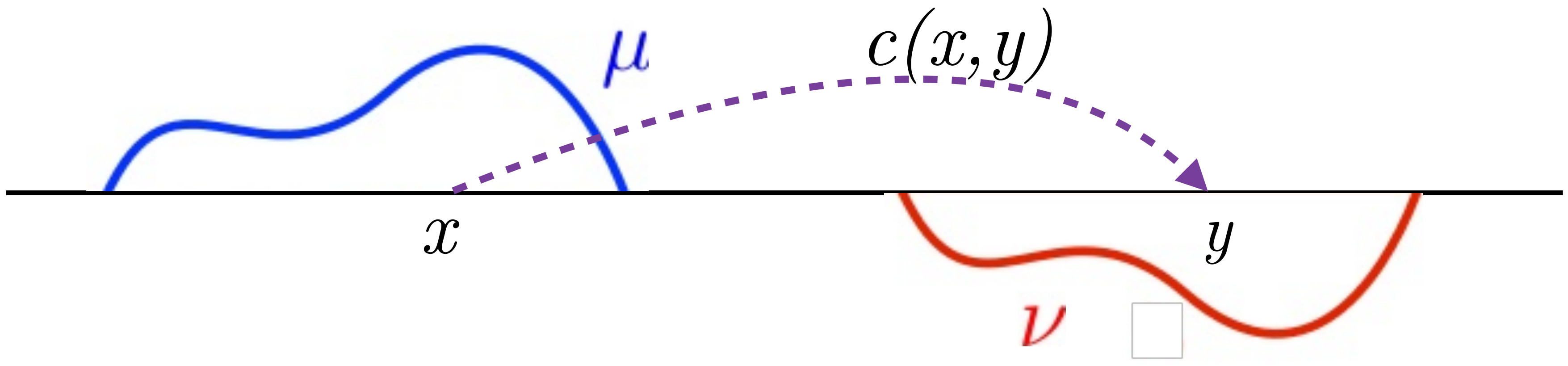
Optimal Transport Distances

Given two probability distributions μ and ν,



Optimal Transport Distances

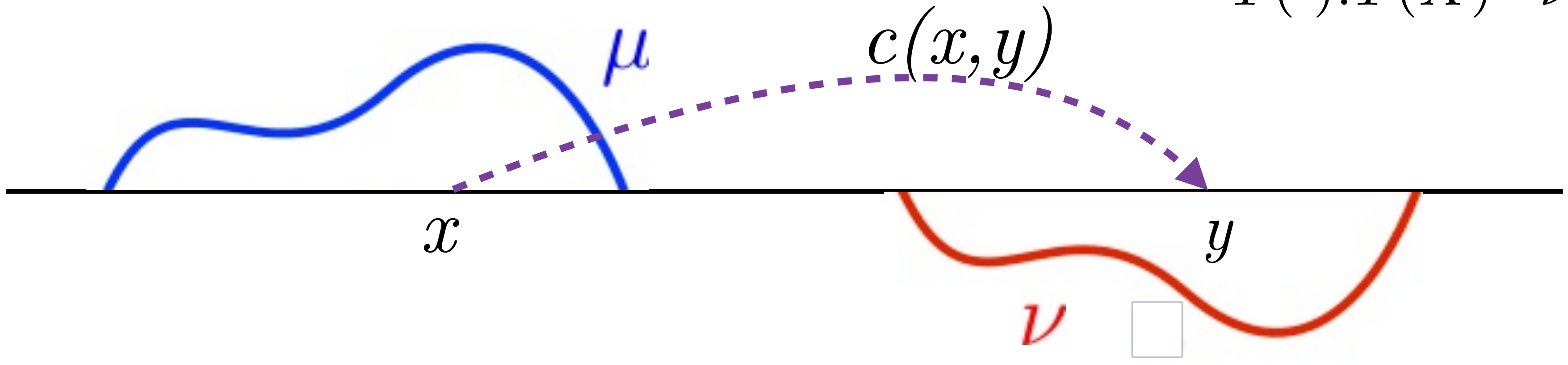
Given two probability distributions μ and ν,



Optimal Transport Distances

Given two probability distributions μ and ν,

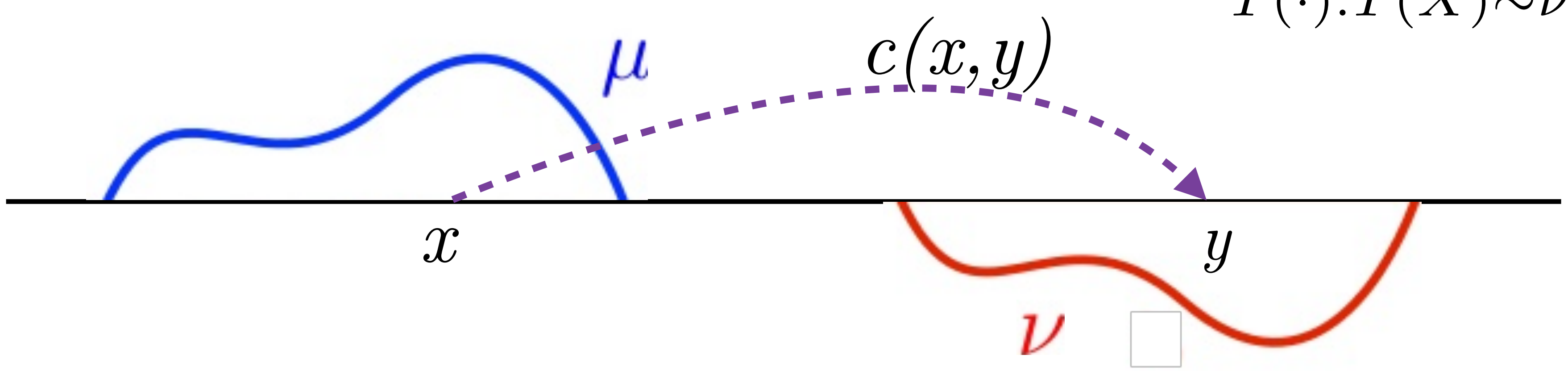
$$\min_{T(\cdot):T(X)\sim\nu} E_\mu\big(c(X,T(X))\big)$$



Optimal Transport Distances

Given two probability distributions μ and ν,

$$\min_{T(\cdot):T(X)\sim\nu} E_\mu\big(c(X, T(X))\big)$$

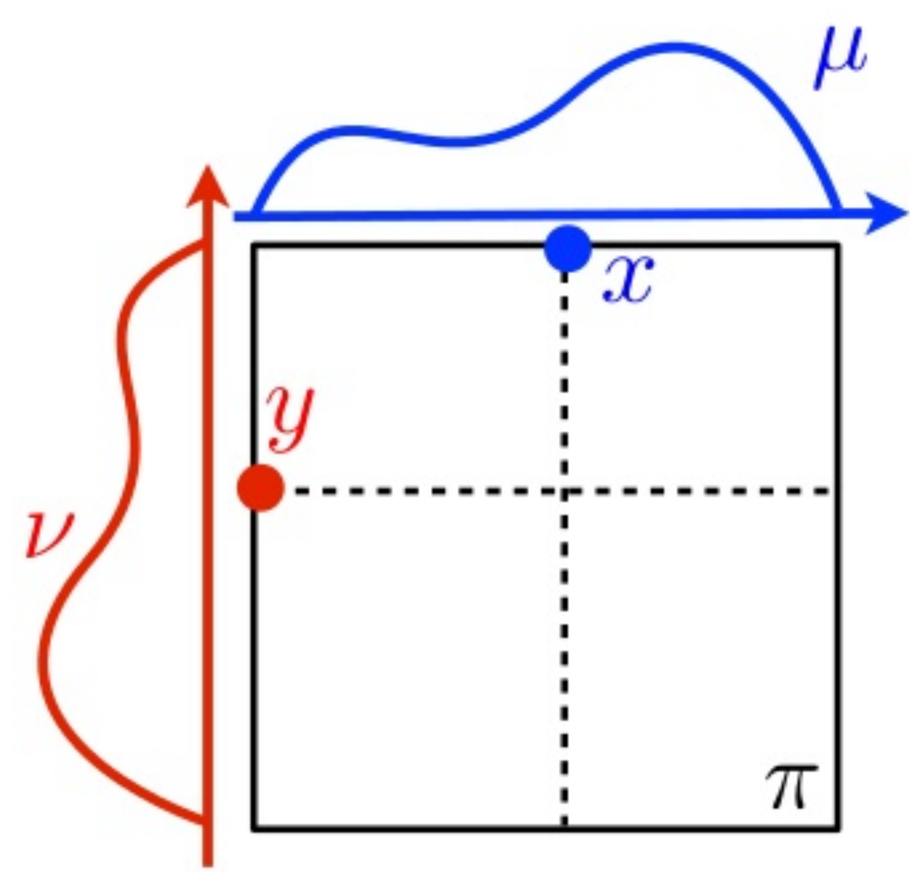


Kantorovich relaxation:

$$D_c(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} E_\pi\big[c(X,Y)\big]$$

X-marginal $=\mu$

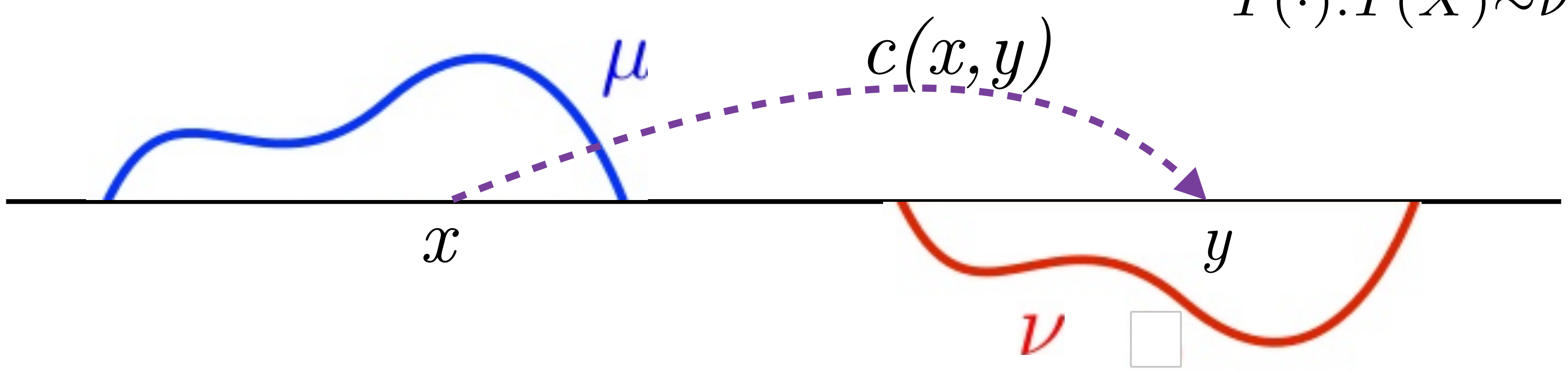
Y-marginal $=\nu$



Optimal Transport Distances

Given two probability distributions μ and ν ,

$$\min_{T(\cdot):T(X)\sim\nu} E_\mu\big(c(X,T(X))\big)$$

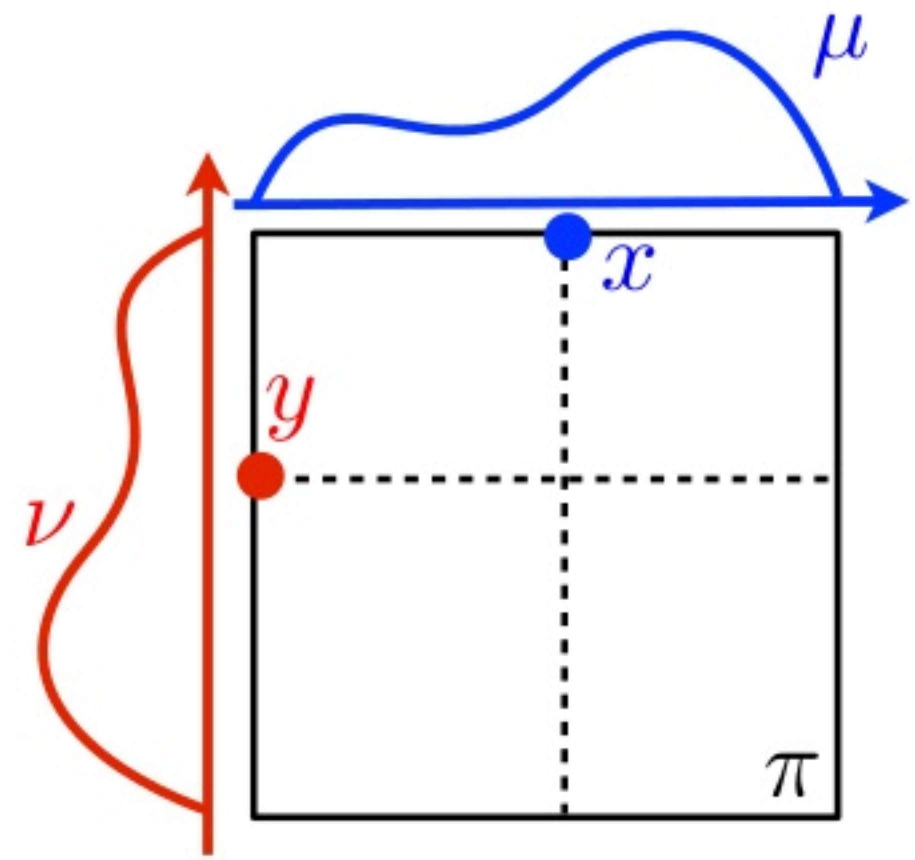


Kantorovich relaxation:

$$D_c(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} E_\pi\big[c(X,Y)\big]$$

X-marginal $=\mu$

Y-marginal $=\nu$



If $c(x,y) = \|x-y\|^p$,

$D_c^{1/p}(\mu,\nu)$ is the Wasserstein distance of order p

Why optimal transport distances?

$$\{P : D_{\mathrm{KL}}(P, P_{ref}) \leq \delta\}$$

Why optimal transport distances?

$$\{P : D_{\mathrm{KL}}(P, P_{ref}) \leq \delta\}$$

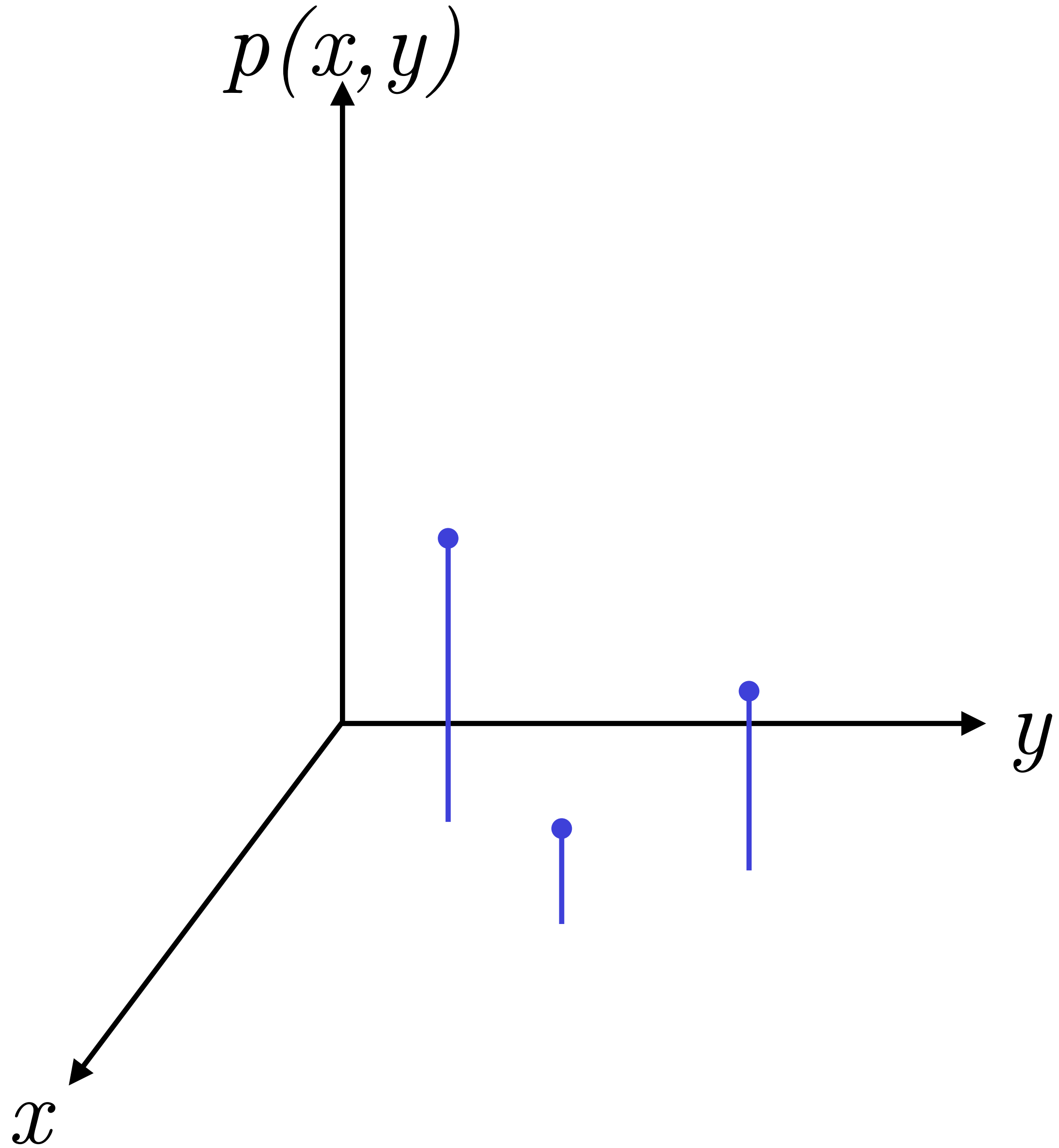
Hansen and Sargent '01, '06
Nilim and El Ghaoui '02, '03
Iyengar '05
Lim and Shanthikumar '04
Lim et al '05, '06
Jain, Lim and Shanthikumar '10
Ben-Tal et al '13
Lam '13, '16, '17
Csiszár and Breuer '13
Jiang and Guan '12
Hu and Hong '13
Wang, Glynn and Ye '14
Glasserman and Xu '14
Bayrakskan and Love '15
Shapiro '15
Duchi, Glynn and Namkoong '16
Dhara, Das and Natarajan '17
Duchi and Namkoong '17

Why optimal transport distances?

$$\{P : D_{\mathrm{KL}}(P, P_{ref}) \leq \delta\}$$

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

Baseline probability distribution *p*

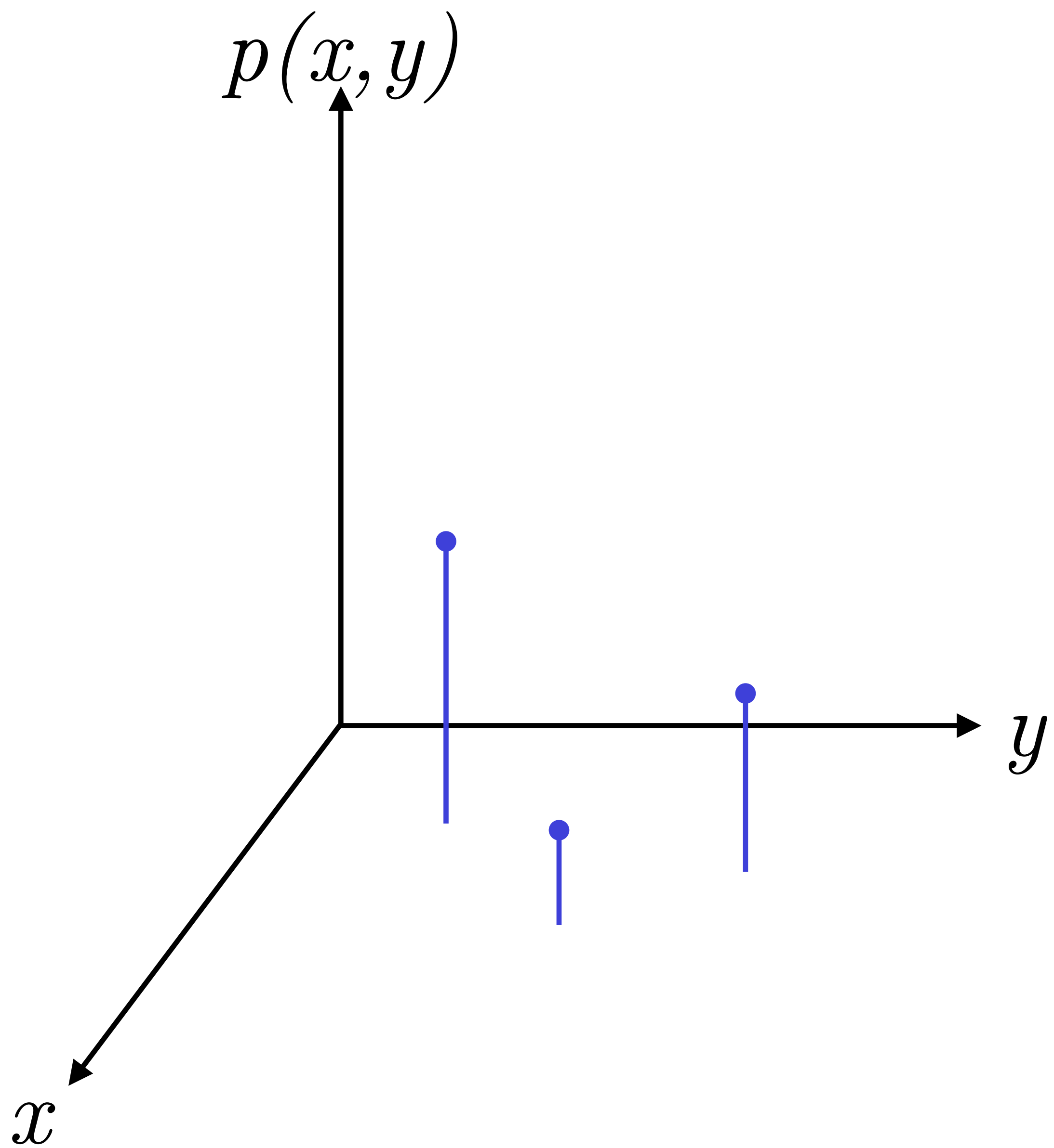


Why optimal transport distances?

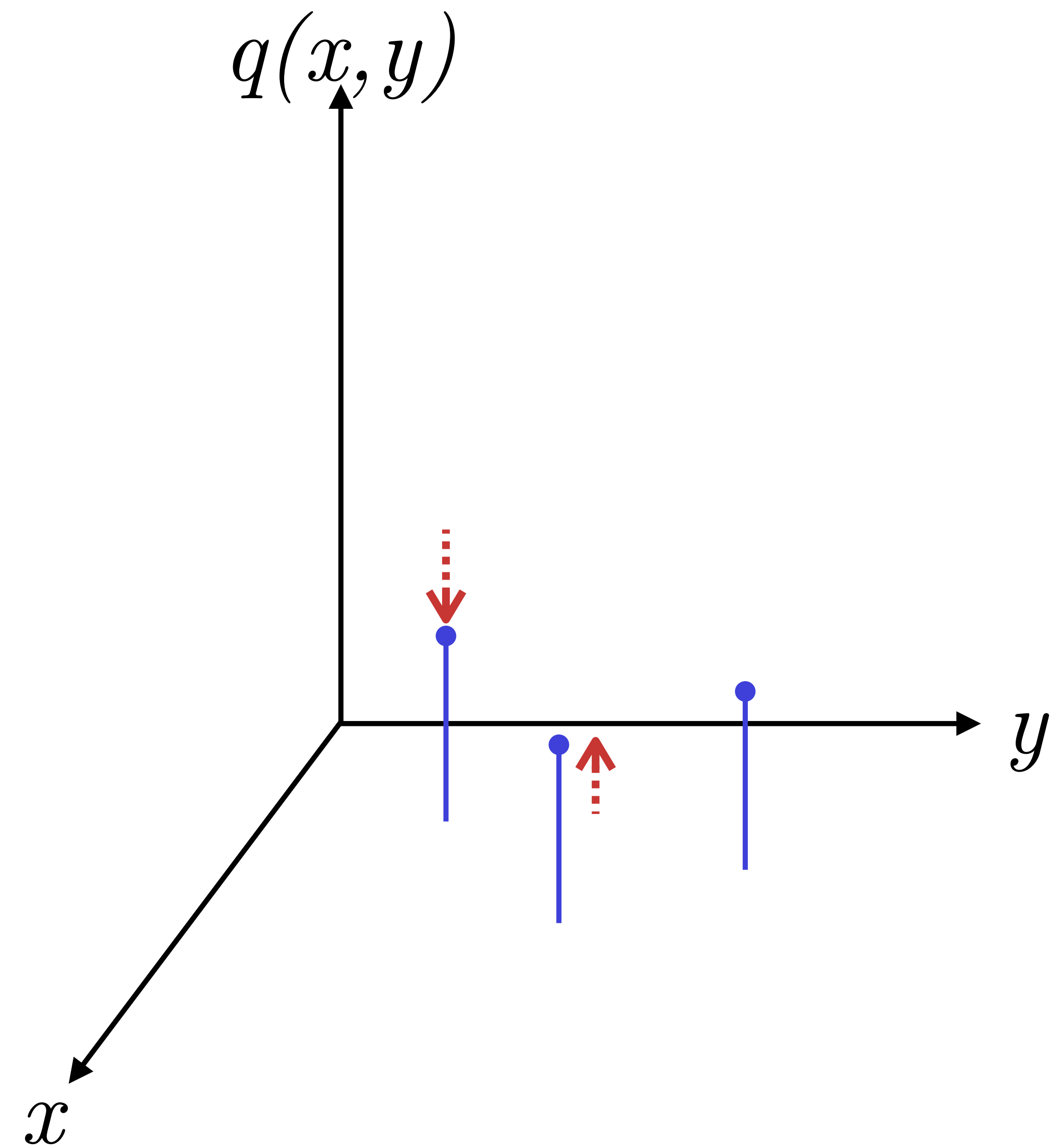
$$\{P : D_{\mathrm{KL}}(P, P_{ref}) \leq \delta\}$$

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

Baseline probability distribution *p*



A KL-neighbor of p

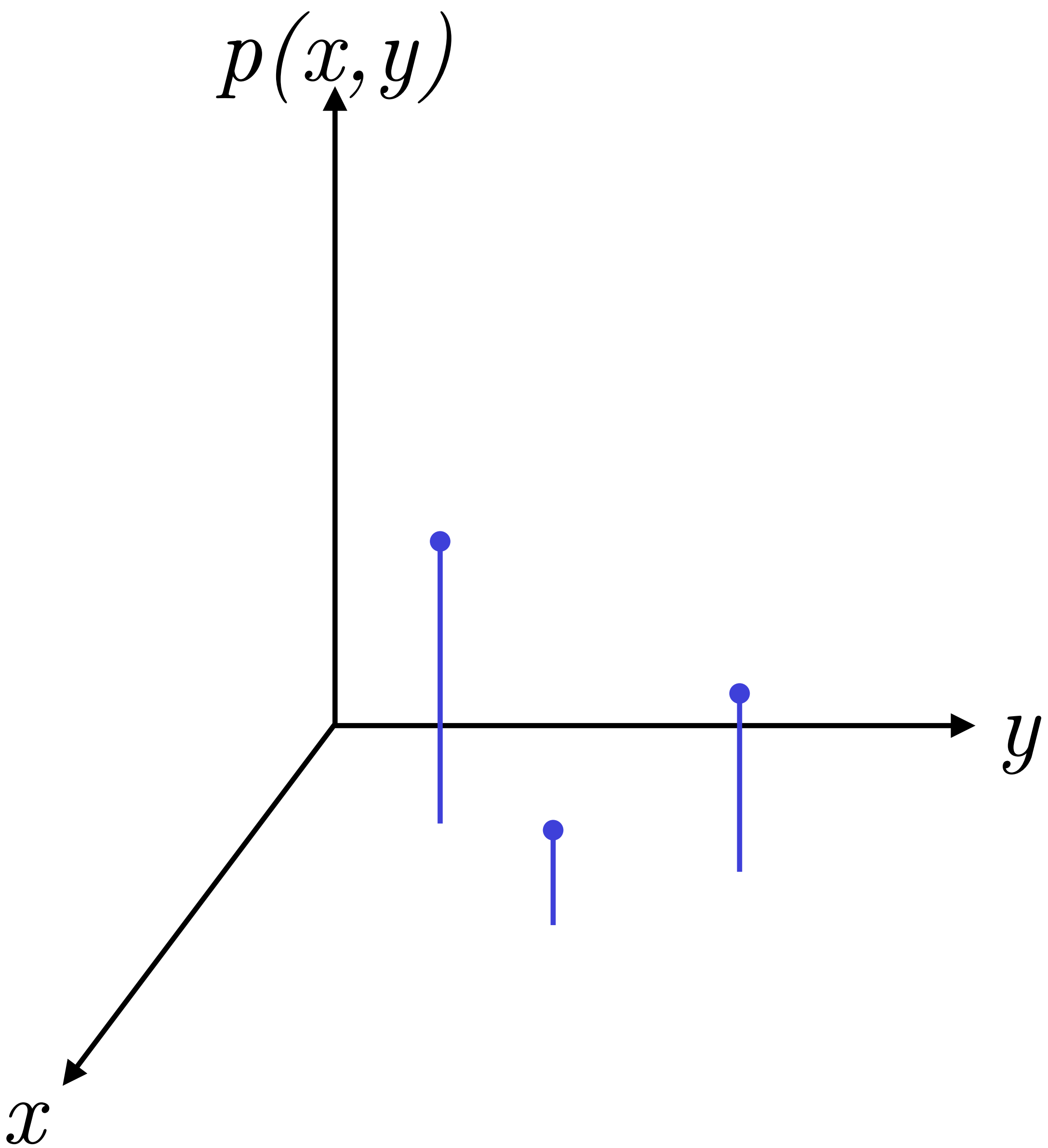


Why optimal transport distances?

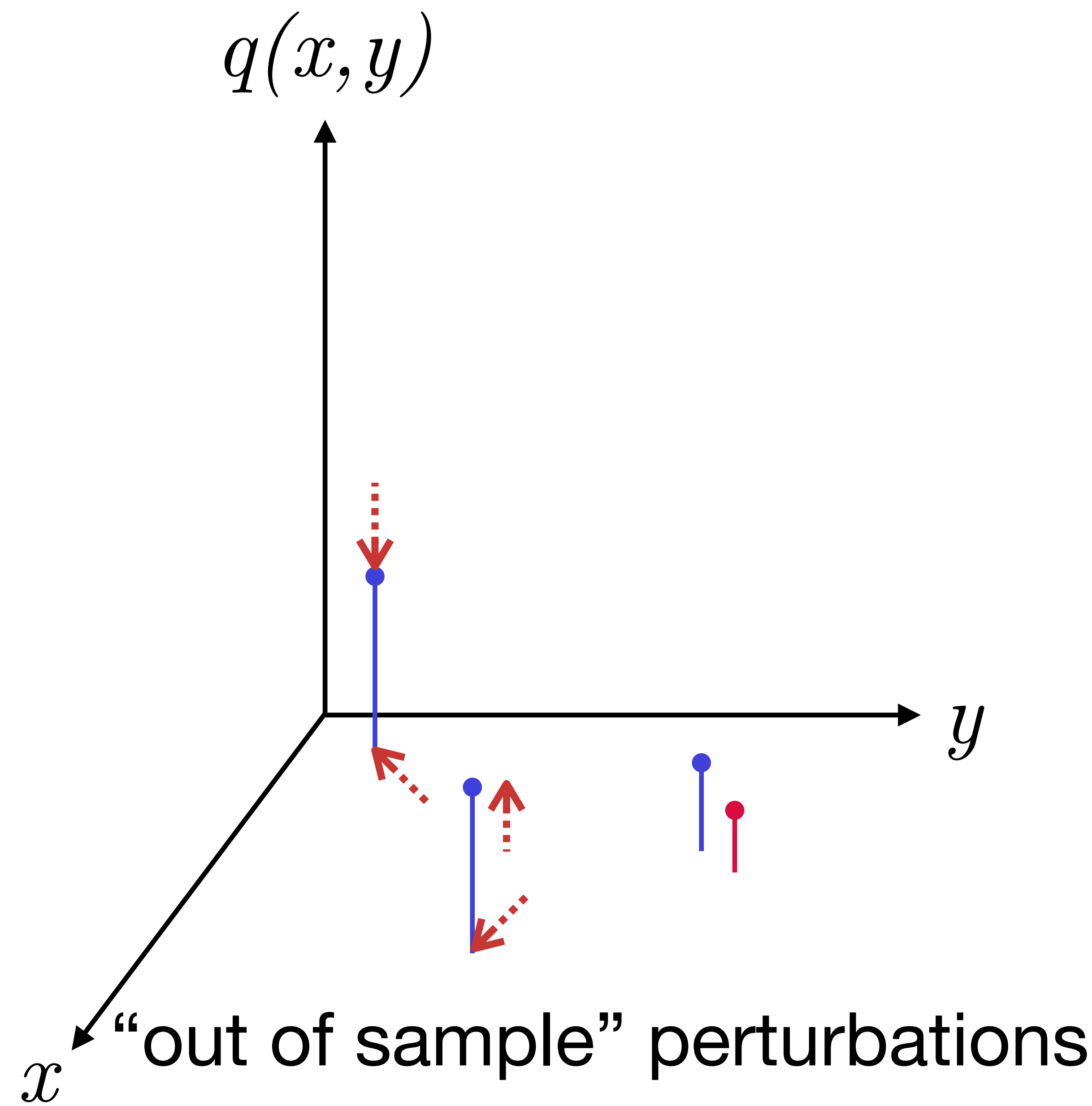
$$\{P : D_{\mathrm{KL}}(P, P_{ref}) \leq \delta\}$$

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

Baseline probability distribution *p*



A Wasserstein neighbor of p



DRO literature that considers optimal transport type distances

Pflug & Wozabal '07
Wozabal '12
Pflug & Pichler '14

DRO literature that considers optimal transport type distances

Pflug & Wozabal ‘07
Wozabal ‘12
Pflug & Pichler ‘14

S-Abadeh, Esfahani & Kuhn ‘15
Lee & Mehrotra ‘15
Esfahani & Kuhn ‘15
Zhao & Guan ‘15
Blanchet & M ‘16
Gao & Kleywegt ‘16
Hanasusanto & Kuhn ‘17
Blanchet, Kang & M ‘17
Blanchet, Kang, Zhang & M ‘17
Luo & Mehrotra ‘17
Lee & Raginsky ‘17
S-Abadeh, Esfahani & Kuhn ‘17
Gao, Chen & Kleywegt ’17

Part I: Recovering well-known regularization based ML estimators as specific examples of DRO

Part I: Recovering well-known regularization based ML estimators as specific examples of DRO

Xu, Caramanis & Mannor (2009a, 2009b)
Bertsimas & Copenhaver (2017)

Distributionally robust linear regression

- Consider fitting a linear regression model

$$Y_i = \beta^T X_i + \varepsilon_i$$

  to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

Distributionally robust linear regression

- Consider fitting a linear regression model
$$Y_i = \beta^T X_i + \varepsilon_i$$
to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

- Optimal least squares finds β that minimizes
$$E_{P_n}\left[(Y - \beta^T X)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta^T X_i\right)^2$$

Distributionally robust linear regression

- Consider fitting a linear regression model
$$Y_i = \beta^T X_i + \varepsilon_i$$
to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

- Optimal least squares finds β that minimizes
$$E_{P_n}\left[(Y - \beta^T X)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta^T X_i\right)^2$$

- DR linear regression: $\min_{\beta} \sup_{P: D_c(P, P_n) \leq \delta} E_P\left[(Y - \beta^T X)^2\right]$

Distributionally robust linear regression

- Consider fitting a linear regression model

$$Y_i = \beta^T X_i + \varepsilon_i$$

  to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

$$D_c(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} E_\pi\left[c(X, Y)\right]$$

- DR linear regression: $\min_{\beta} \sup_{P: D_c(P, P_n) \leq \delta} E_P\left[(Y - \beta^T X)^2\right]$

Distributionally robust linear regression

- Consider fitting a linear regression model

$$Y_i = \beta^T X_i + \varepsilon_i$$

to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

$$D_c(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} E_\pi\left[c(X,Y)\right]$$

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

Theorem (Blanchet, Kang & M '16)

Suppose $c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ \infty & \text{if } y \neq y' \end{cases}$

Distributionally robust linear regression

- Consider fitting a linear regression model

$$Y_i = \beta^T X_i + \varepsilon_i$$

to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

$$D_c(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} E_\pi \left[c(X, Y) \right]$$

- DR linear regression: $\min_\beta \sup_{P : D_c(P, P_n) \leq \delta} E_P \left[(Y - \beta^T X)^2 \right]$

Theorem (Blanchet, Kang & M '16)

Suppose $c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ \infty & \text{if } y \neq y' \end{cases}$. Then

DR-linear regression estimator $= \arg\min_\beta \left\{ \sqrt{\text{MSE}_n(\beta)} + \sqrt{\delta} \|\beta\|_p \right\}$

Distributionally robust linear regression

- Consider fitting a linear regression model

$$Y_i = \beta^T X_i + \varepsilon_i$$

to data points $(X_1, Y_1), \ldots, (X_n, Y_n)$

Image source: r-bloggers.com

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

Theorem (Blanchet, Kang & M '16)

Suppose $c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ \infty & \text{if } y \neq y' \end{cases}$. Then

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

$1/p + 1/q = 1$

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

Theorem (Blanchet, Kang & M '16)

Suppose $c\big((x,y),(x',y')\big) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ \infty & \text{if } y \neq y' \end{cases}$. Then $1/p + 1/q = 1$

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

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

Theorem (Blanchet, Kang & M '16)

Suppose $c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q^2 & \text{if } y = y', \\ \infty & \text{if } y \neq y' \end{cases}$. Then $1/p + 1/q = 1$

DR-linear regression estimator $= \arg\min_{\beta} \left\{ \sqrt{\text{MSE}_n(\beta)} + \sqrt{\delta} \|\beta\|_p \right\}$

- DR logistic regression: $\min_{\beta} \sup_{P: D_c(P,P_n) \le \delta} E_P\left[\text{Logistic Loss}(X, Y; \beta)\right]$

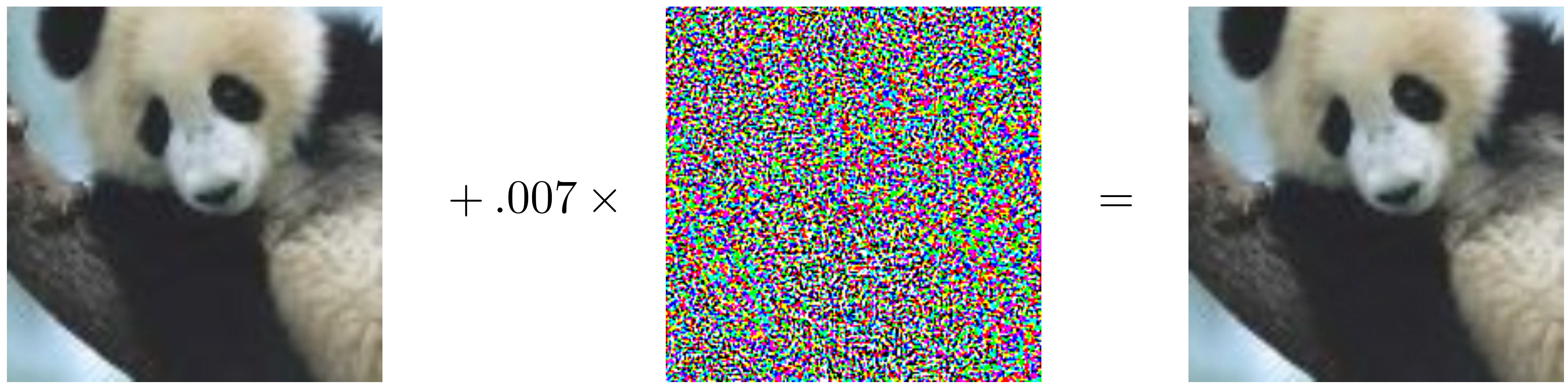
Theorem (Blanchet, Kang & M '16)

Suppose $c((x,y),(x',y')) = \begin{cases} \|x - x'\|_q & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$. Then

DR-logistic regression estimator

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

Image from [Szegedy et al 2015]



- DR logistic regression: $\min_{\beta} \sup_{P: D_c(P,P_n) \le \delta} E_P\left[\text{Logistic Loss}(X,Y;\beta)\right]$

Theorem (Blanchet, Kang & M '16)

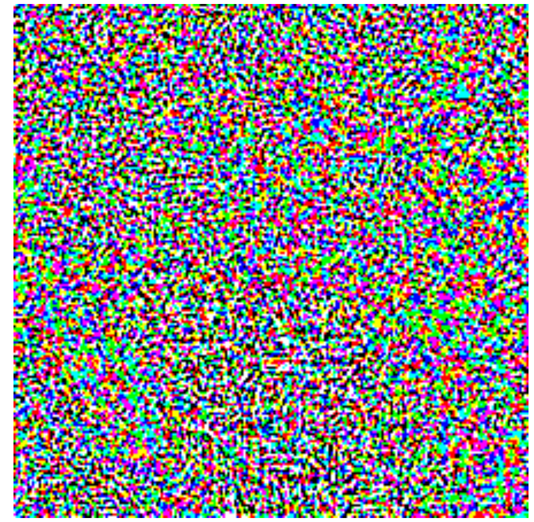
Suppose $c\big((x,y),(x',y')\big) = \begin{cases} \|x-x'\|_q & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$. Then

DR-logistic regression estimator

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

[Szegedy et al 2015]

$+ .007 \times$



$=$

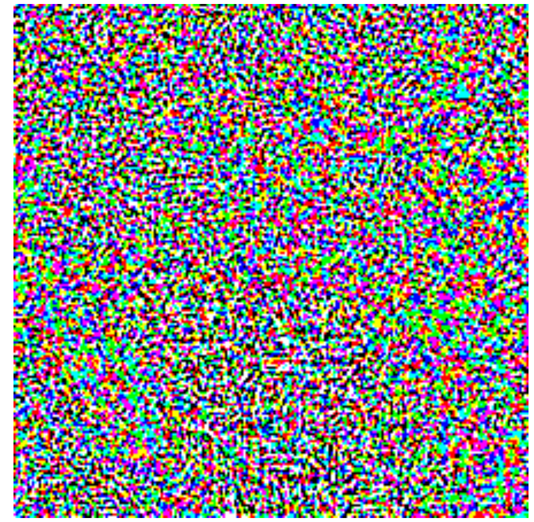
$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

[Szegedy et al 2015]

$+ .007 \times$



$=$

$\boldsymbol{x}$

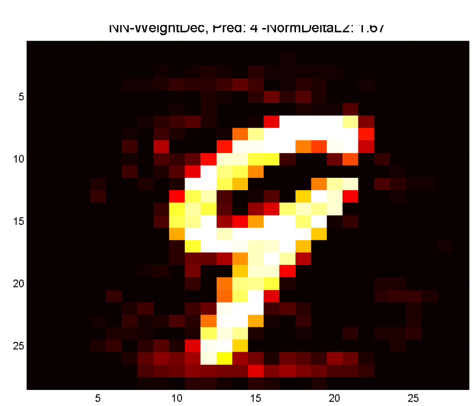
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

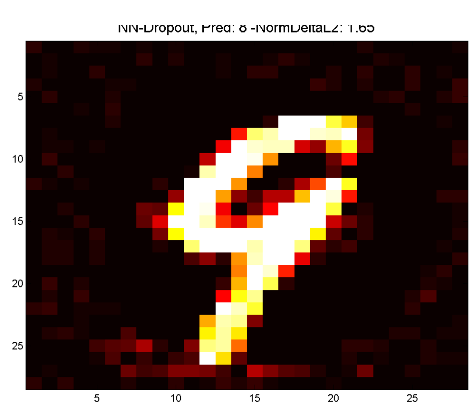
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence



NN-WD, Pred:4, $\|\delta\|_2 = 1.7$



NN-DO, Pred:8, $\|\delta\|_2 = 1.7$

S-Abadeh, Esfahani & Kuhn (2015)

S-Abadeh, Esfahani & Kuhn (2015)

Blanchet, Kang & M ‘16

Blanchet, Kang, Zhang & M ‘17

S-Abadeh, Esfahani & Kuhn ’17

Gao, Chen & Kleywegt ‘17

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

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

Theorem (Blanchet, Kang & M '16)

Suppose $c\big((x,y),(x',y')\big) = \begin{cases} \|x-x'\|_q^2 & \text{if } y=y', \\ \infty & \text{if } y\neq y' \end{cases}$. Then $1/p+1/q=1$

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

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

Duality Theorem (Blanchet & M '16)

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

Esfahani & Kuhn '15, Zhao & Guan '15
Gao & Kleywegt '16

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

Duality Theorem (Blanchet & M '16)

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

Esfahani & Kuhn '15, Zhao & Guan '15
Gao & Kleywegt '16

General assumption:

cost c is lower semicontinuous

cost can be infinity

f is upper semicontinuous

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

Duality Theorem (Blanchet & M ‘16)

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

Esfahani & Kuhn ’15, Zhao & Guan ‘15
Gao & Kleywegt ’16

General assumption:

- cost c is lower semicontinuous
- cost can be infinity
- f is upper semicontinuous

Applications in

- risk analysis
- data driven optimization
- stochastic control
- machine learning, ...

- DR-linear regression (with q -norm cost) $= \ell_p$-regularized linear regression
  - q=1 case exactly recovers $\sqrt{\text{Lasso}}$
  - q=2 case recovers ridge regression
- DR-logistic regression (with q -norm cost) $= \ell_p$-reg. logistic regression

- DR-linear regression (with q -norm cost) $= \ell_p$-regularized linear regression
  - $q{=}1$ case exactly recovers $\sqrt{\text{Lasso}}$
  - $q{=}2$ case recovers ridge regression
- DR-logistic regression (with q -norm cost) $= \ell_p$-reg. logistic regression

DRO with optimal transport costs recovers many other regularized estimators....

- DR-hinge loss minimization = Support Vector Machines
- DR-quantile regression (with q -norm cost) $= \ell_p$-reg. quantile regression
- Group lasso, LAD-Lasso
- Generalized adaptive ridge regression

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P [\ell(X; \beta)] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

- A number of popular ML algorithms can be exactly recast as particular examples of (OT-DRO) ✔
- Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

$$\inf_{\beta} \sup_{P: D(P, P_n) \le \delta} E_P [\ell(X; \beta)] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

→ A number of popular ML algorithms can be exactly recast as particular examples of (OT-DRO) ✔

→ Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

[Esfahani & Kuhn '15]

[Kuhn & Hanasusanto '17]

[Luo & Mehrotra '17]

[Sinha, Namkoong & Duchi '17]

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \quad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

→ A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO) ✔

→ Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

convex $\ell(\beta^T X)$ or $\max_{i=1,\ldots,K} \ell_i(\beta^T X)$

$$\inf_{\beta} \sup_{P: D(P,P_n) \le \delta} E_P\left[\ell(X;\beta)\right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

→ A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO) ✔

→ Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets?

convex $\ell(\beta^T X)$ or $\max_{i=1,\ldots,K} \ell_i(\beta^T X)$

Linear, Logistic, Poisson regression...
Multi-task learning
Kernel-based algorithms
Multinomial logit models

Utility maximization
Newsvendor models

Part II: Fast iterative schemes for optimal transport DRO

(work in progress)

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P : D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

→

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \le \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

→

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$

$$\nabla f_1, \quad \nabla f_2, \quad \ldots, \quad \nabla f_{n-1}, \quad \nabla f_n$$

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$
- SGD scheme: $\begin{bmatrix} \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \beta_k \\ \lambda_k \end{bmatrix} - \alpha_k \begin{bmatrix} \partial f_I / \partial \beta \\ \partial f_I / \partial \lambda \end{bmatrix} (\beta_k, \lambda_k), \quad k = 1, 2, \ldots,$

$$\nabla f_1, \quad \nabla f_2, \quad \ldots, \quad \nabla f_{n-1}, \quad \nabla f_n$$

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$
- SGD scheme: $\begin{bmatrix} \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \beta_k \\ \lambda_k \end{bmatrix} - \alpha_k \begin{bmatrix} \partial f_I / \partial \beta \\ \partial f_I / \partial \lambda \end{bmatrix} (\beta_k, \lambda_k), \quad k = 1, 2, \ldots,$
- After T iterations, error $= O(1/T)$ if F is strongly convex

  error $= O(1/\sqrt{T})$ if F is convex

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \le \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A(x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$

$$f_i(\beta, \lambda) := \sup_{\gamma_i \in \mathbb{R}} \left\{ \ell\left(Y_i, \beta^T X_i + \gamma_i \sqrt{\delta} \beta^T A^{-1} \beta \right) - \lambda \sqrt{\delta} \left(\gamma_i^2 \beta^T A^{-1} \beta - 1 \right) \right\}.$$

First order oracle information can be evaluated just with function evaluations of $\ell(\cdot)$ and $\ell'(\cdot)$, which is what ERM also requires

ERM:

$$\min_{\beta \in B} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P,P_n) \leq \delta} E_P\left[\ell(Y_i, \beta^T X_i)\right]$$

- Take $c(x,y) = (x-y)^T A (x-y)$
- Applying the Duality theorem, $\inf_{\beta,\lambda}\left\{F(\beta,\lambda) := \frac{1}{n}\sum_{i=1}^{n} f_i(\beta,\lambda)\right\}$

$$f_i(\beta,\lambda) := \sup_{\gamma_i \in \mathbb{R}} \left\{\ell\left(Y_i, \beta^T X_i + \gamma_i \sqrt{\delta}\beta^T A^{-1}\beta\right) - \lambda\sqrt{\delta}\left(\gamma_i^2 \beta^T A^{-1}\beta - 1\right)\right\}.$$

$$\frac{\partial f_i}{\partial \lambda} = -\sqrt{\delta}\left(\gamma_i^2 \beta^T A^{-1}\beta - 1\right) \qquad \frac{\partial f_i}{\partial \beta} = \ell'(Y_i, \beta^T \tilde{X}_i)\tilde{X}_i$$

First order oracle information can be evaluated just with function evaluations of $\ell(\cdot)$ and $\ell'(\cdot)$, which is what ERM also requires

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A(x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$
- SGD scheme: $\begin{bmatrix} \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \beta_k \\ \lambda_k \end{bmatrix} - \alpha_k \begin{bmatrix} \partial f_I / \partial \beta \\ \partial f_I / \partial \lambda \end{bmatrix} (\beta_k, \lambda_k), \quad k = 1, 2, \ldots,$

	ERM	DRO
Per-iteration complexity	$O(d)$	$O(Ld)$
# Iterations		
Complexity		

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \le \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$

Theorem

Suppose $\ell(X; \beta) = \max_{i=1,\ldots,K} \ell_i(\beta^T X)$, where $\ell_i \in C^2$ are locally strongly convex.

Then for all $\delta < \delta_0$, the function F is strongly convex with parameter $= c\sqrt{\delta}$.

.

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

→

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x,y) = (x-y)^T A (x-y)$
- Applying the Duality theorem, $\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$

Theorem

Suppose $\ell(X; \beta) = \max_{i=1,\ldots,K} \ell_i(\beta^T X)$, where $\ell_i \in C^2$ are locally strongly convex.

Then for all $\delta < \delta_0$, the function F is strongly convex with parameter $= c\sqrt{\delta}$.

Further, F is strongly convex in β as long as $\ell_i \in C^2$ are convex.

ERM:

$$\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta^T X_i)$$

OT-DRO:

$$\min_{\beta \in B} \sup_{P: D_c(P, P_n) \le \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$

- Take $c(x, y) = (x - y)^T A (x - y)$
- Applying the Duality theorem, $$\inf_{\beta, \lambda} \left\{ F(\beta, \lambda) := \frac{1}{n} \sum_{i=1}^{n} f_i(\beta, \lambda) \right\}$$
- SGD scheme: $$\begin{bmatrix} \beta_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} \beta_k \\ \lambda_k \end{bmatrix} - \alpha_k \begin{bmatrix} \partial f_I / \partial \beta \\ \partial f_I / \partial \lambda \end{bmatrix} (\beta_k, \lambda_k), \quad k = 1, 2, \dots,$$

	ERM	DRO
Per-iteration complexity	$O(d)$	$O(Ld)$
# Iterations	$O(\varepsilon^{-2})$	$O(\varepsilon^{-1}\delta^{-1/2})$
Complexity	$O(d\varepsilon^{-2})$	$O(Ld\varepsilon^{-1}\delta^{-2})$

when strong convexity holds

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

- → A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO) ✔
- → Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets? ✔

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P\left[\ell(X;\beta)\right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

- → A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO) ✔
- → Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets? ✔
- → How do we specify the parameters for the ambiguity model?

$$\inf_{\beta} \sup_{P:D(P,P_n)\leq\delta} E_P\left[\ell(X;\beta)\right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

As a flexible & scalable approach towards data-driven optimization under uncertainty

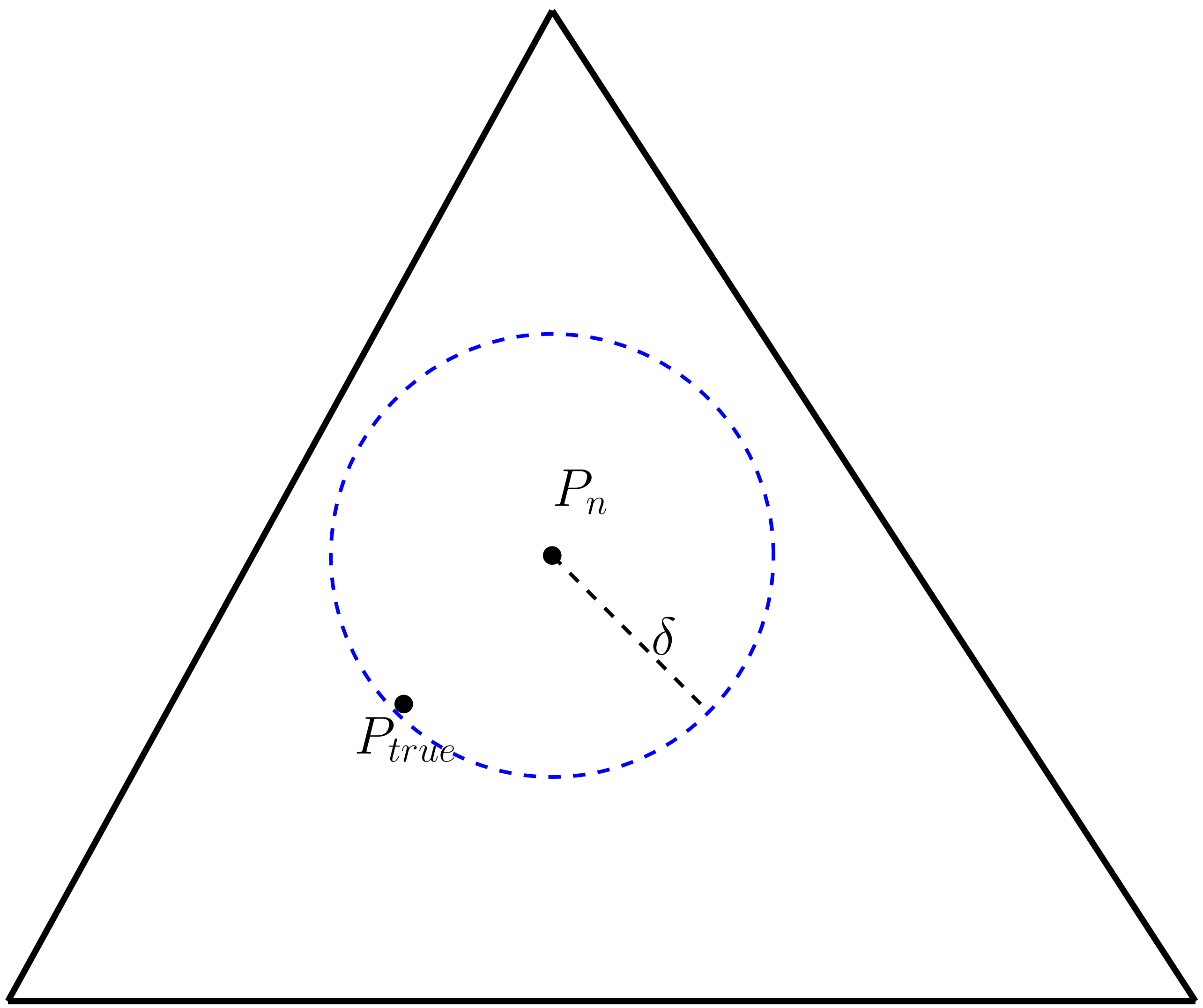
- → A number of popular ML algorithms that employ regularization can be exactly recast as particular examples of (OT-DRO) ✔
- → Can we utilise (OT-DRO) for larger class of models with the ability to handle large data sets? ✔
- → How do we specify the parameters for the ambiguity model?
  - → choosing the radius
  - → utilising data to inform the geometry of the ambiguity region

Part III: Specifying parameters of the optimal transport neighborhood

Specifying radius of the ambiguity models

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

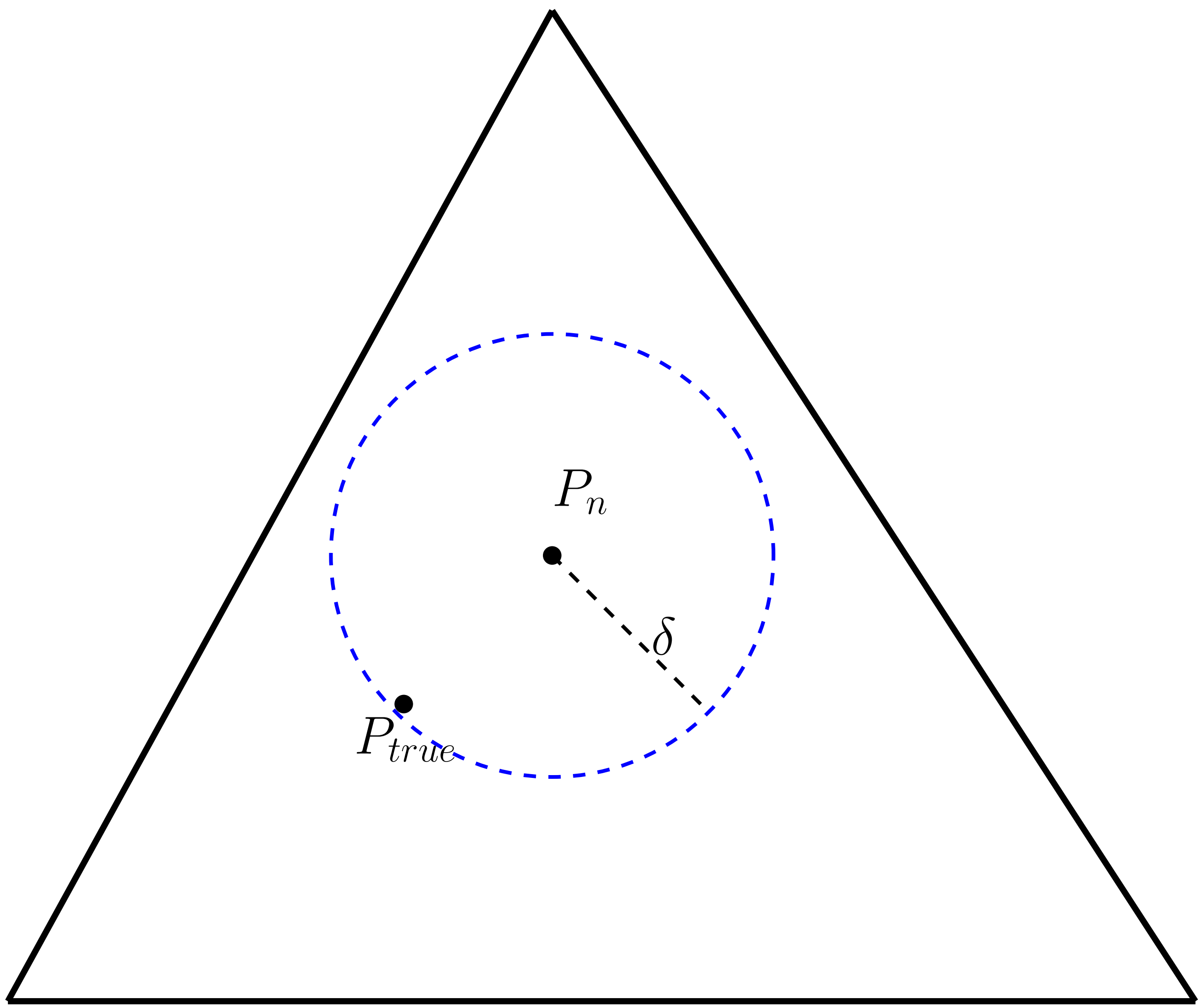
$$P\left(D_c(P_{true}, P_n) \leq \delta\right) \geq 1 - \varepsilon$$



Specifying radius of the ambiguity models

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

$$P\left(D_c(P_{true}, P_n) \le \delta\right) \ge 1 - \varepsilon$$



Concentration inequalities by Fournier & Guillin (2015)

S-Abadeh, Esfahani & Kuhn ‘15, Lee and Mehrotra ’15, Gao and Kleywegt ‘16

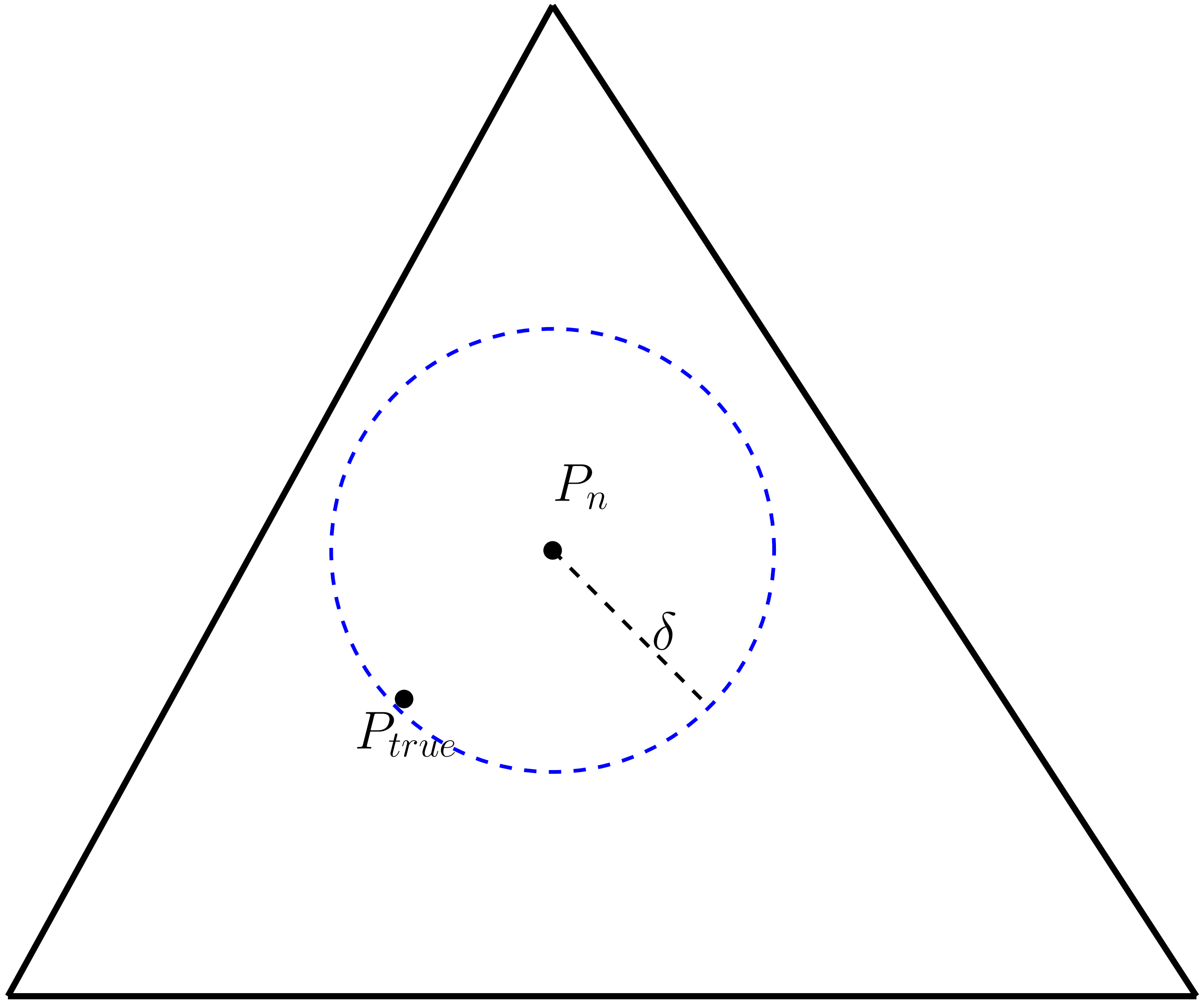
Specifying radius of the ambiguity models

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

Given P,

$\beta_{(P)} :=$ optimal β satisfying

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



Specifying radius of the ambiguity models

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

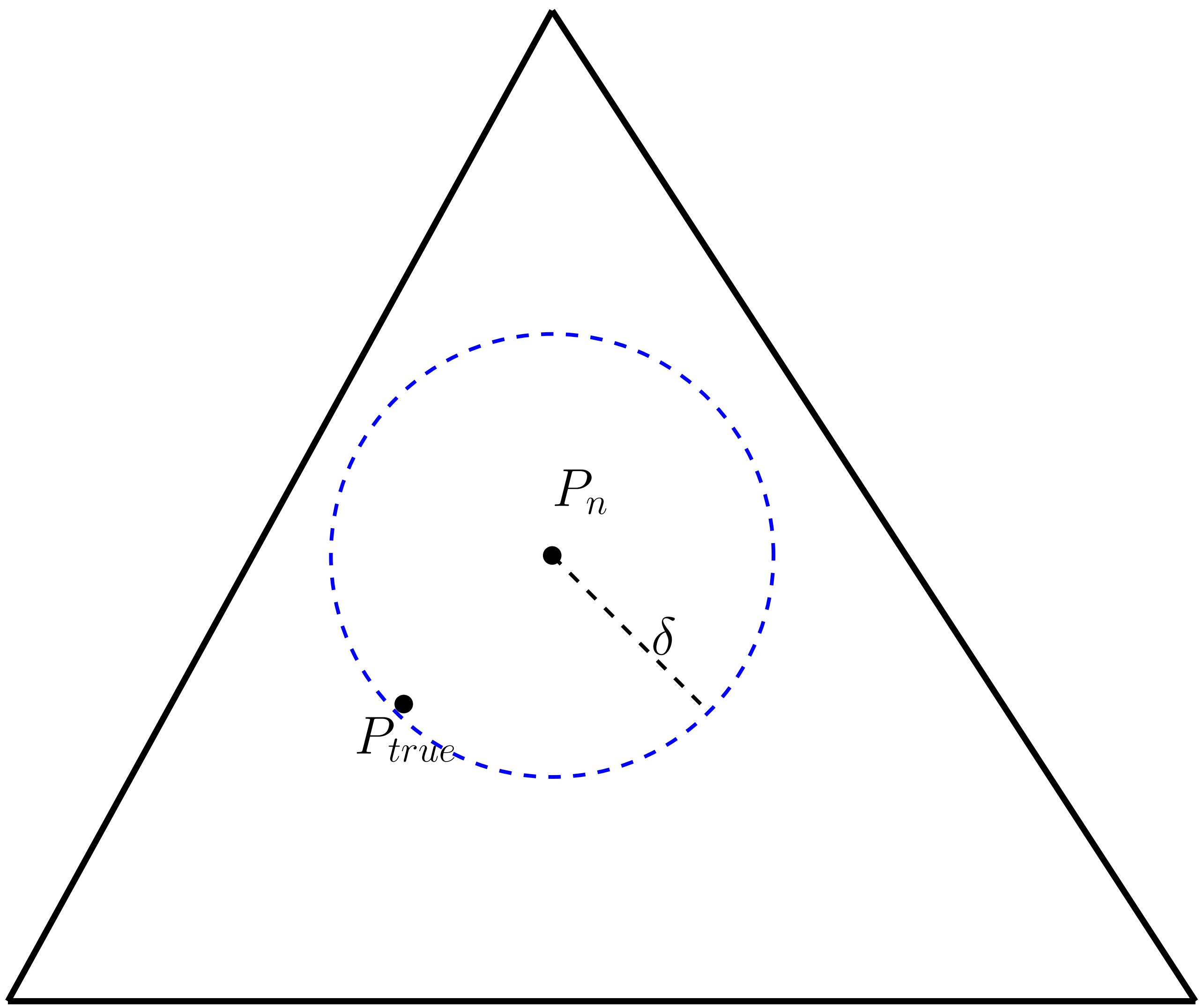
Given P,

$\beta_{(P)} :=$ optimal β satisfying

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

β_* is the optimal β satisfying

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



Specifying radius of the ambiguity models

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

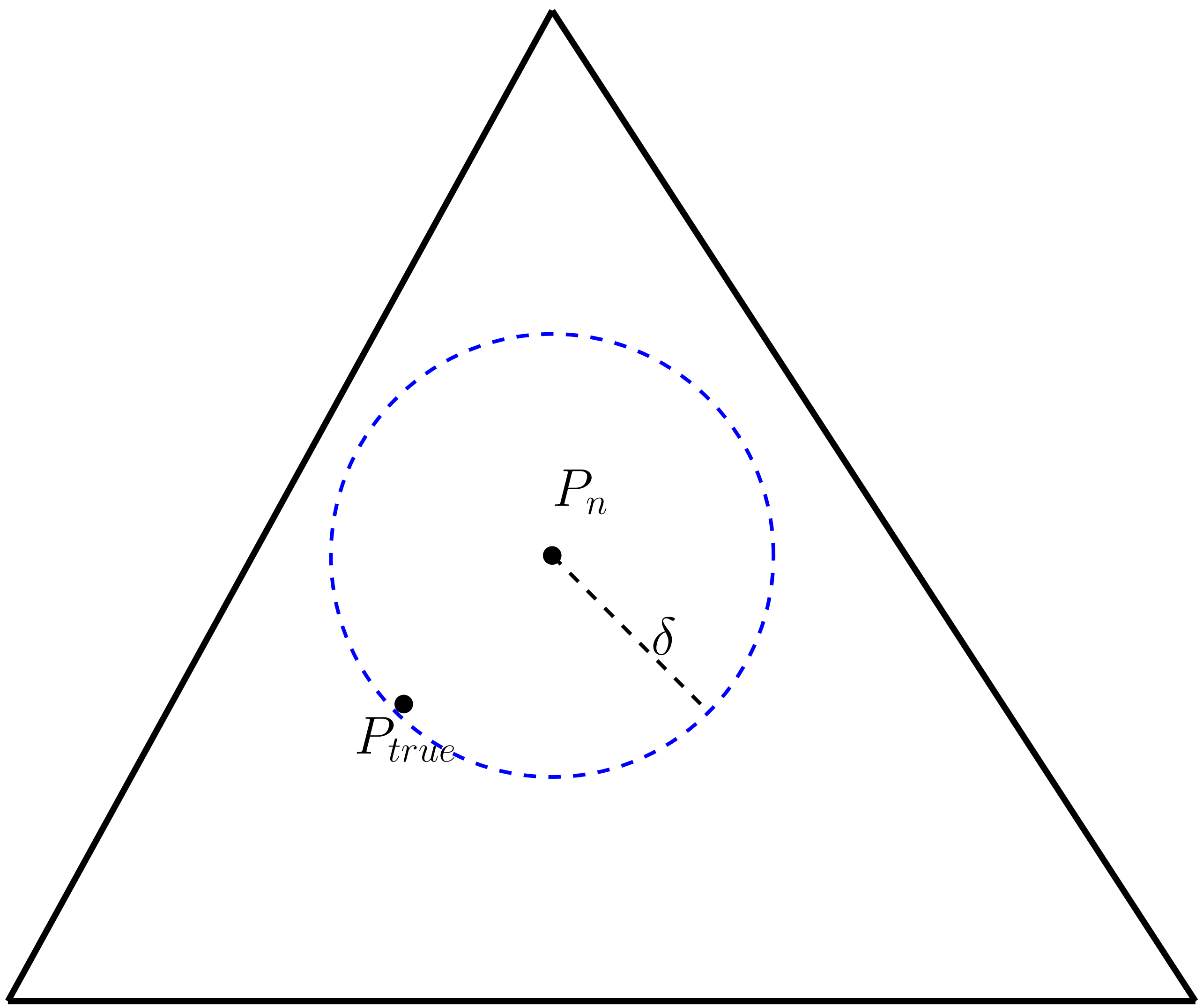
Given P,

$\beta_{(P)} :=$ optimal β satisfying

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

β_* is the optimal β satisfying

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

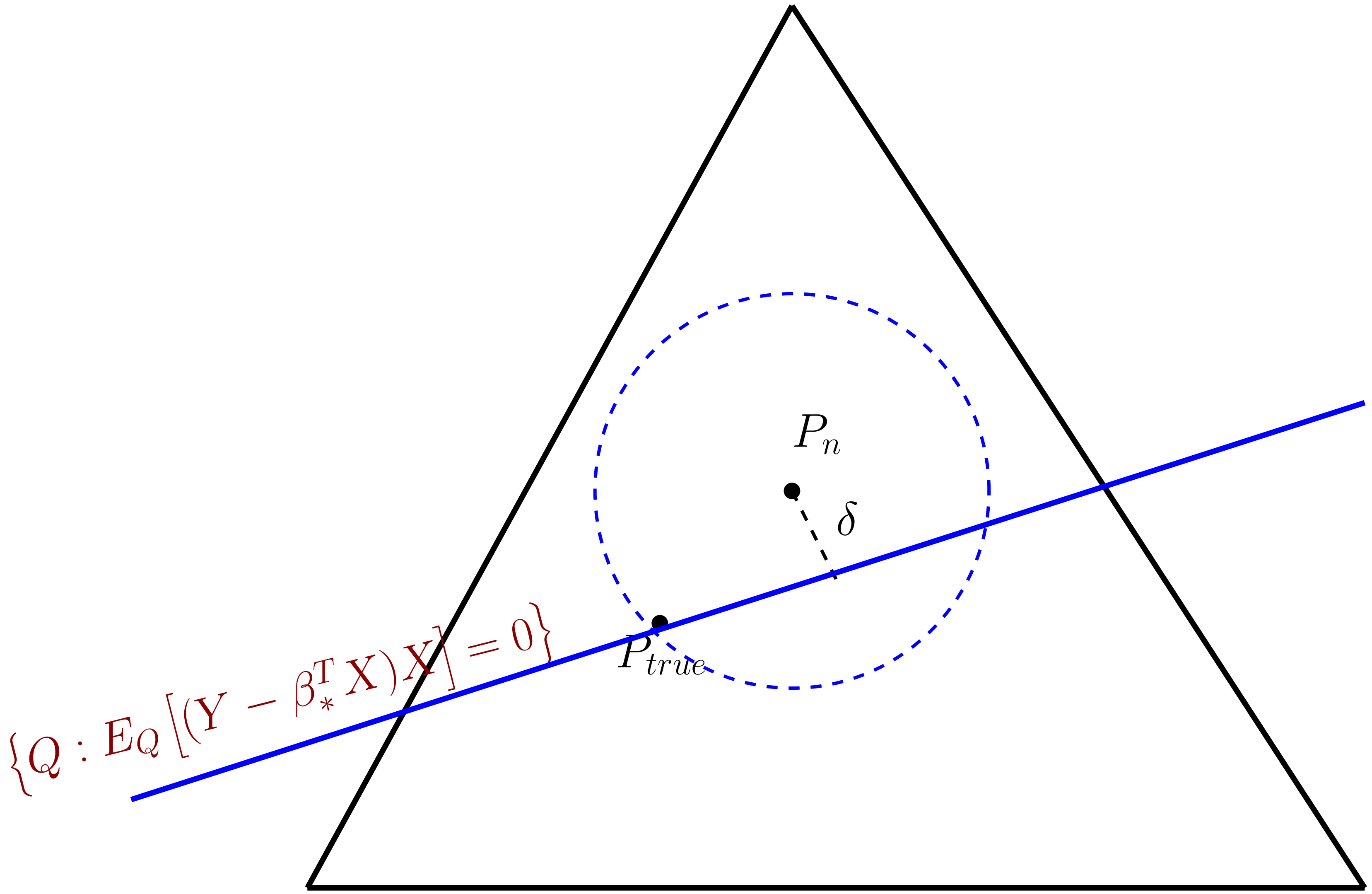


Plausible β's:

Criteria for optimal selection: $\beta_* \in \left\{ \beta_{(P)} : D_c(P, P_n) \le \delta \right\}$

Specifying radius of the ambiguity models

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



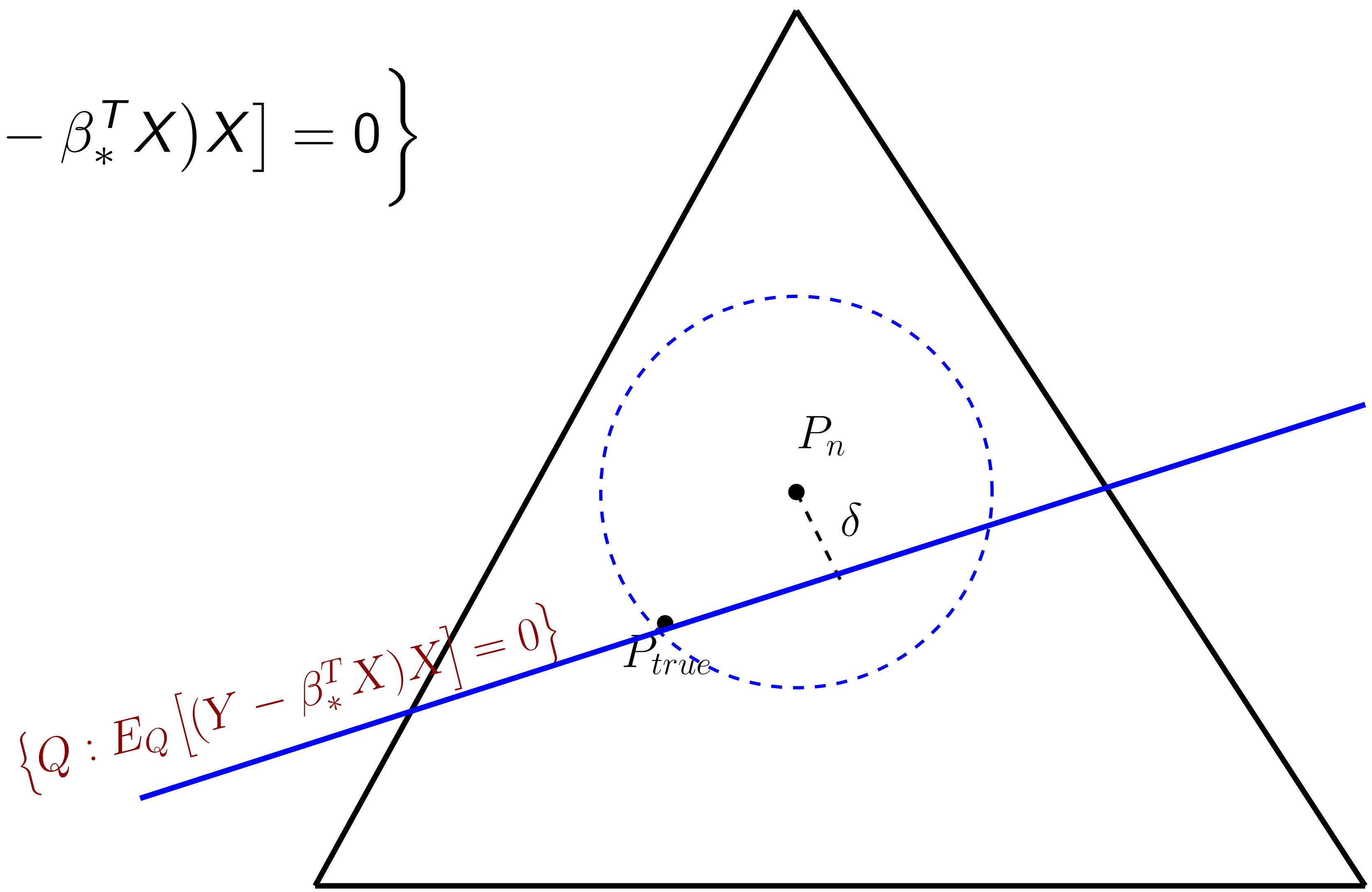
Plausible β's:

Criteria for optimal selection: $\beta_* \in \{\beta_{(P)} : D_c(P, P_n) \le \delta\}$

Specifying radius of the ambiguity models

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

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



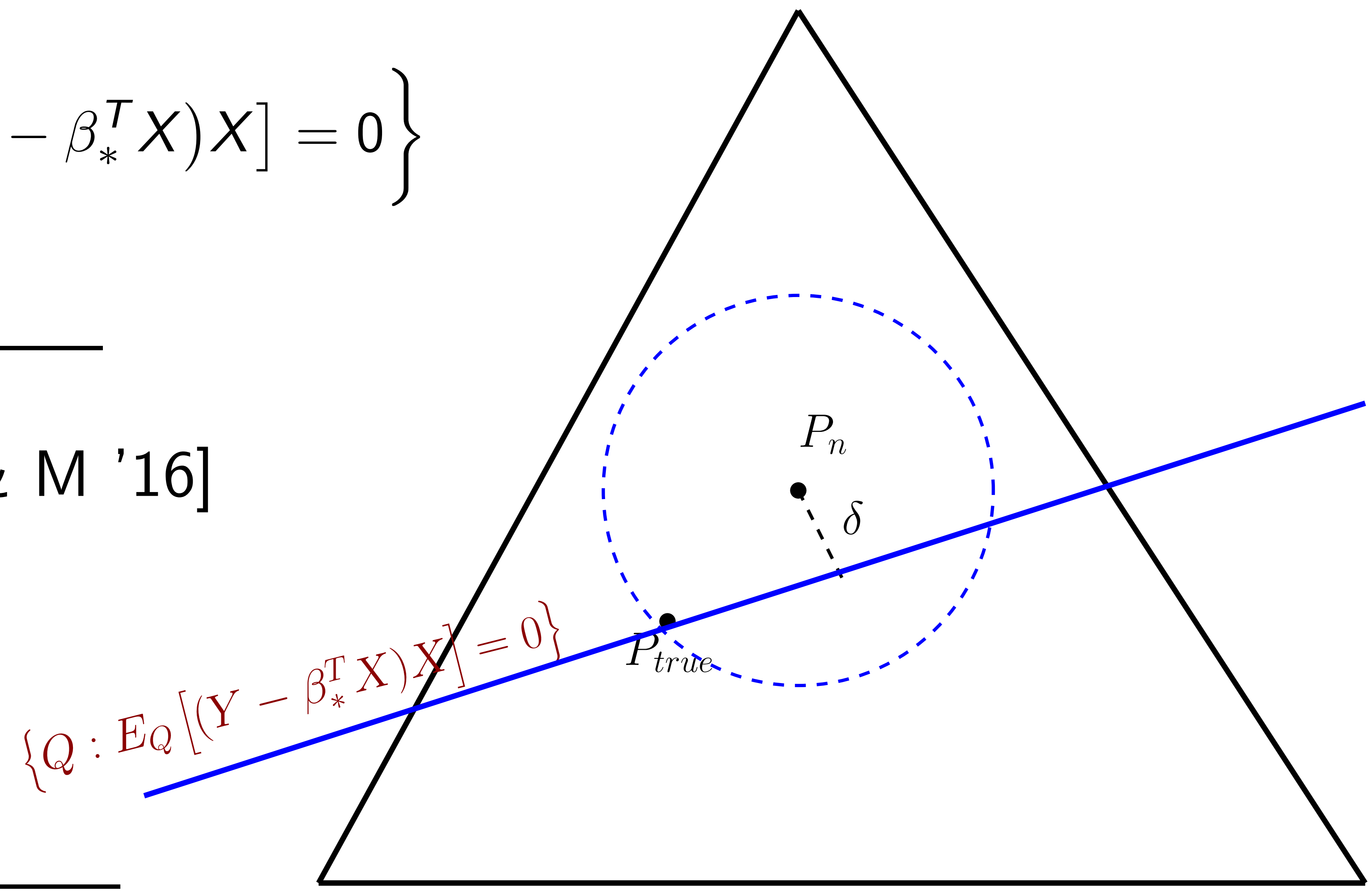
Plausible β's:

Criteria for optimal selection: $\beta_* \in \left\{ \beta_{(P)} : D_c(P, P_n) \leq \delta \right\}$

Specifying radius of the ambiguity models

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

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



Theorem: [Blanchet, Kang & M '16]

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

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

Plausible β's:

Criteria for optimal selection: $\beta_* \in \left\{\beta_{(P)} : D_c(P, P_n) \leq \delta\right\}$

Specifying radius of the ambiguity models

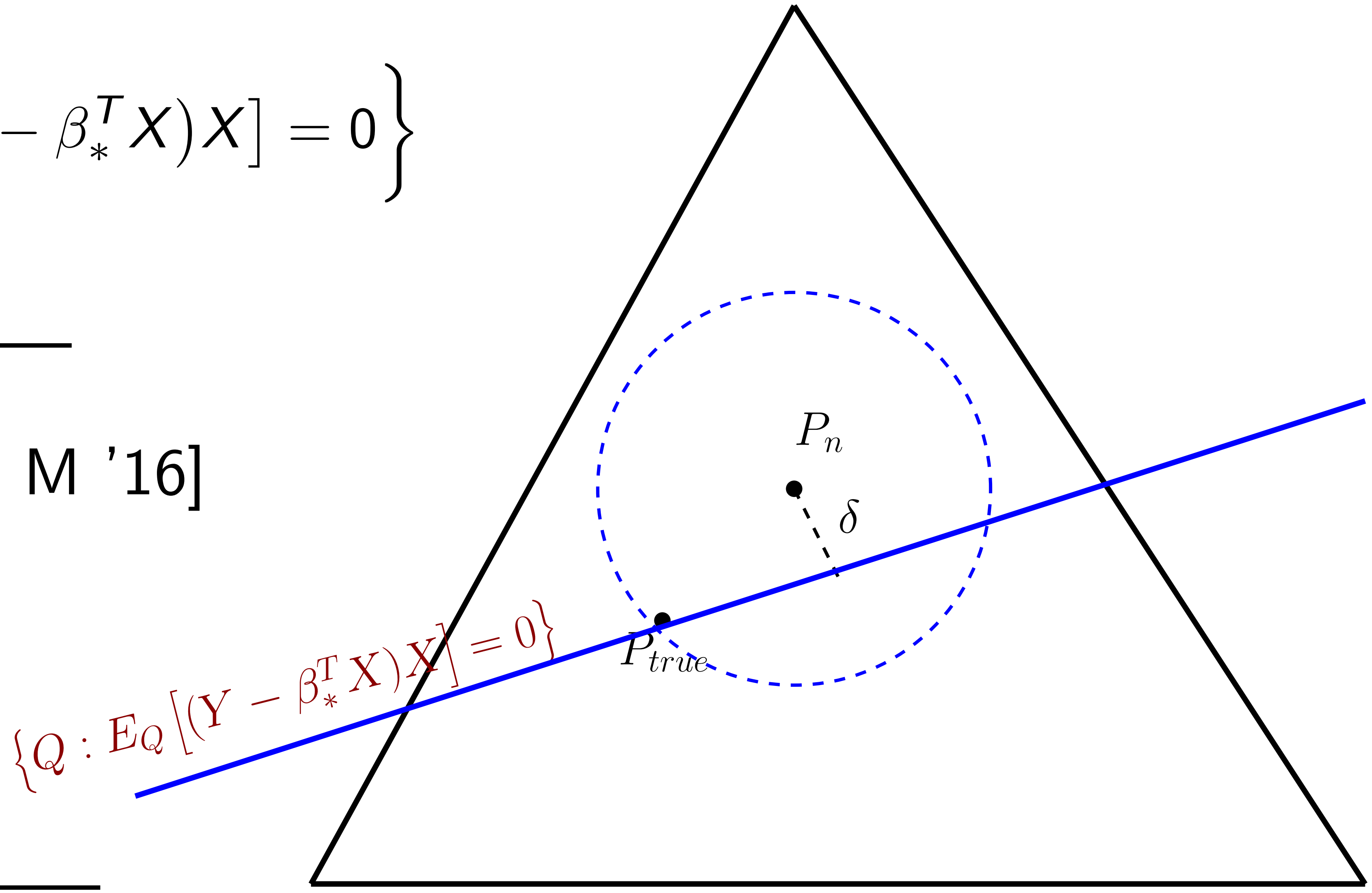
DR linear regression: $\min_{\beta \in \mathbb{R}^d} \max_{P: D_c(P,P_n) \le \delta} E_P\left[\left(Y - \beta^T X\right)^2\right]$

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

Theorem: [Blanchet, Kang & M '16]

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

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

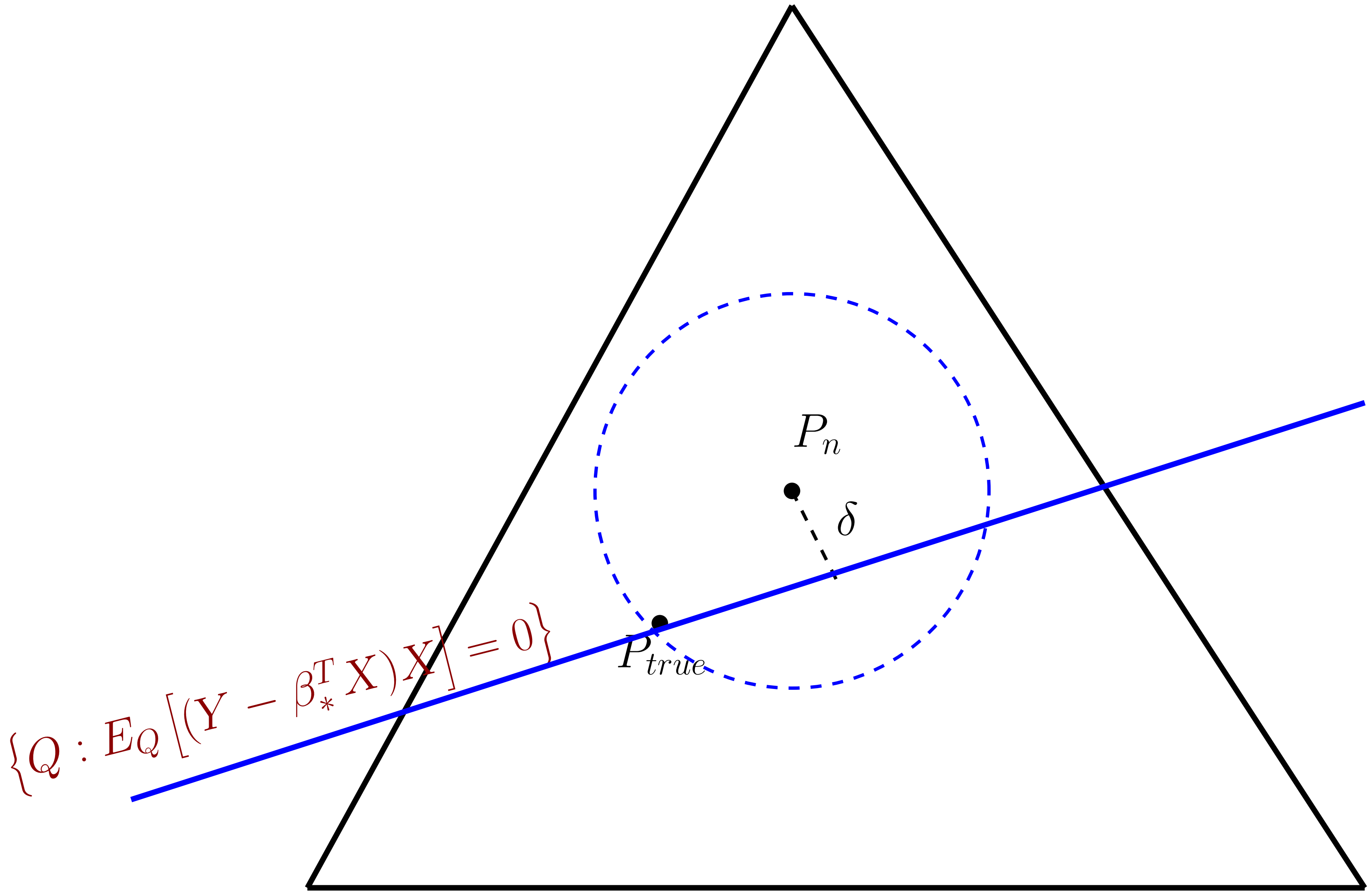


Choose $\delta = \frac{\eta_\alpha}{n}$ where η_α is such that $P\left\{\bar{R} \le \eta_\alpha\right\} = 1 - \alpha$.

Specifying radius of the ambiguity models

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

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



Specifying radius of the ambiguity models

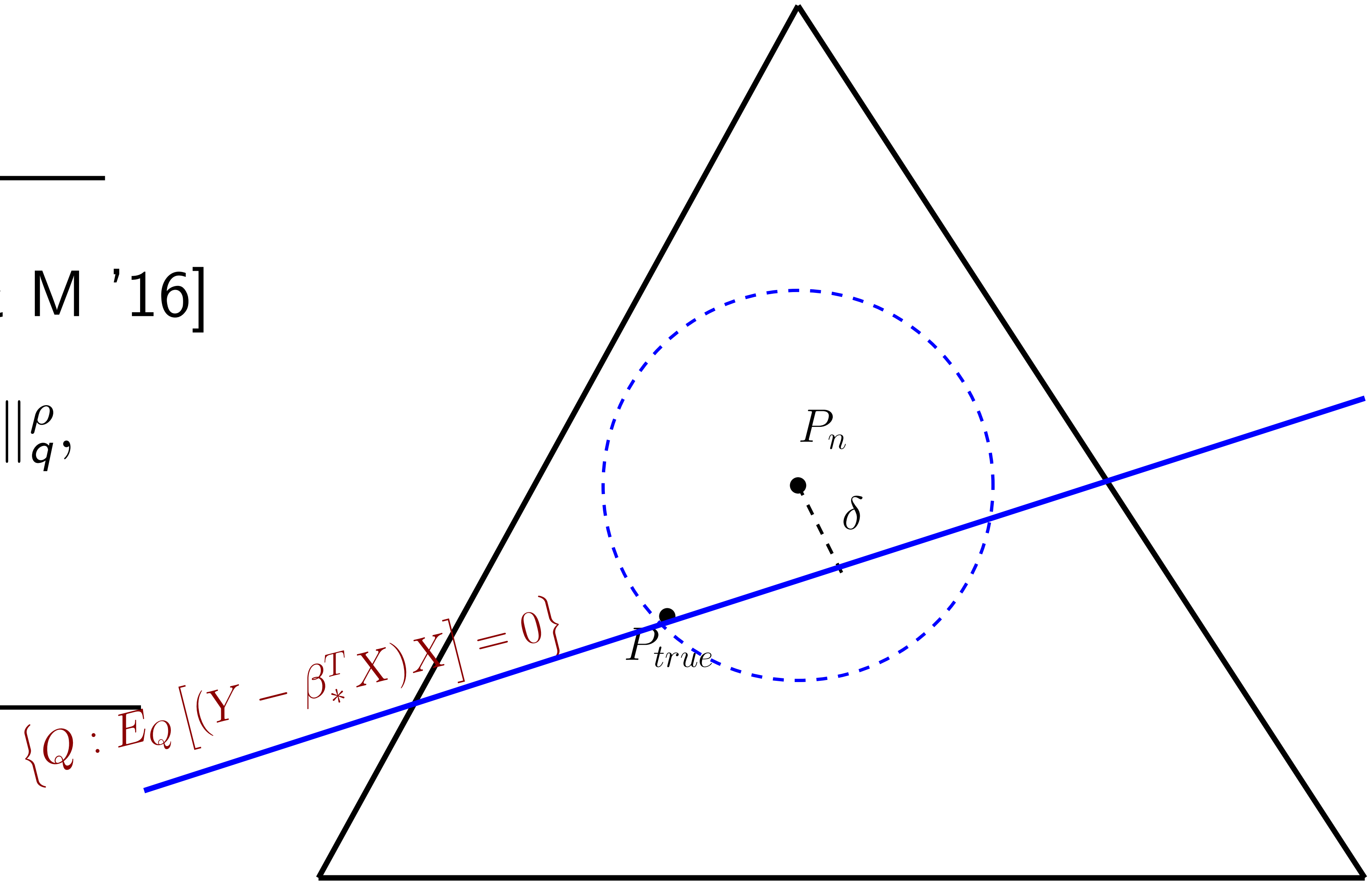
`Optimality condition:` $E\left[h(W;\beta_*)\right] = \mathbf{0}$

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

Theorem: [Blanchet, Kang & M '16]

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

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



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

Specifying radius of the ambiguity models

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

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

Theorem: [Blanchet, Kang & M '16]

$$c(u,v) = \|u - v\|_q^{\rho},$$

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

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

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

Specifying radius of the ambiguity models

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

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

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

Theorem: [Blanchet, Kang & M '16]

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

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

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

Specifying radius of the ambiguity models

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

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

Theorem: [Blanchet, Kang & M '16]

$$c(u,v) = \|u - v\|_q^{\rho},$$

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

ℓ_p−lin reg: $\rho = 2$

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

ℓ_p−log reg: $\rho = 1$

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

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

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

Specifying radius of the ambiguity models

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

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

Theorem: [Blanchet, Kang & M '16]

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

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

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

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

Application to machine learning: No cross-validation!

Application 1: `DR linear regression`

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

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

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

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

Application 2: `DR logistic regression`

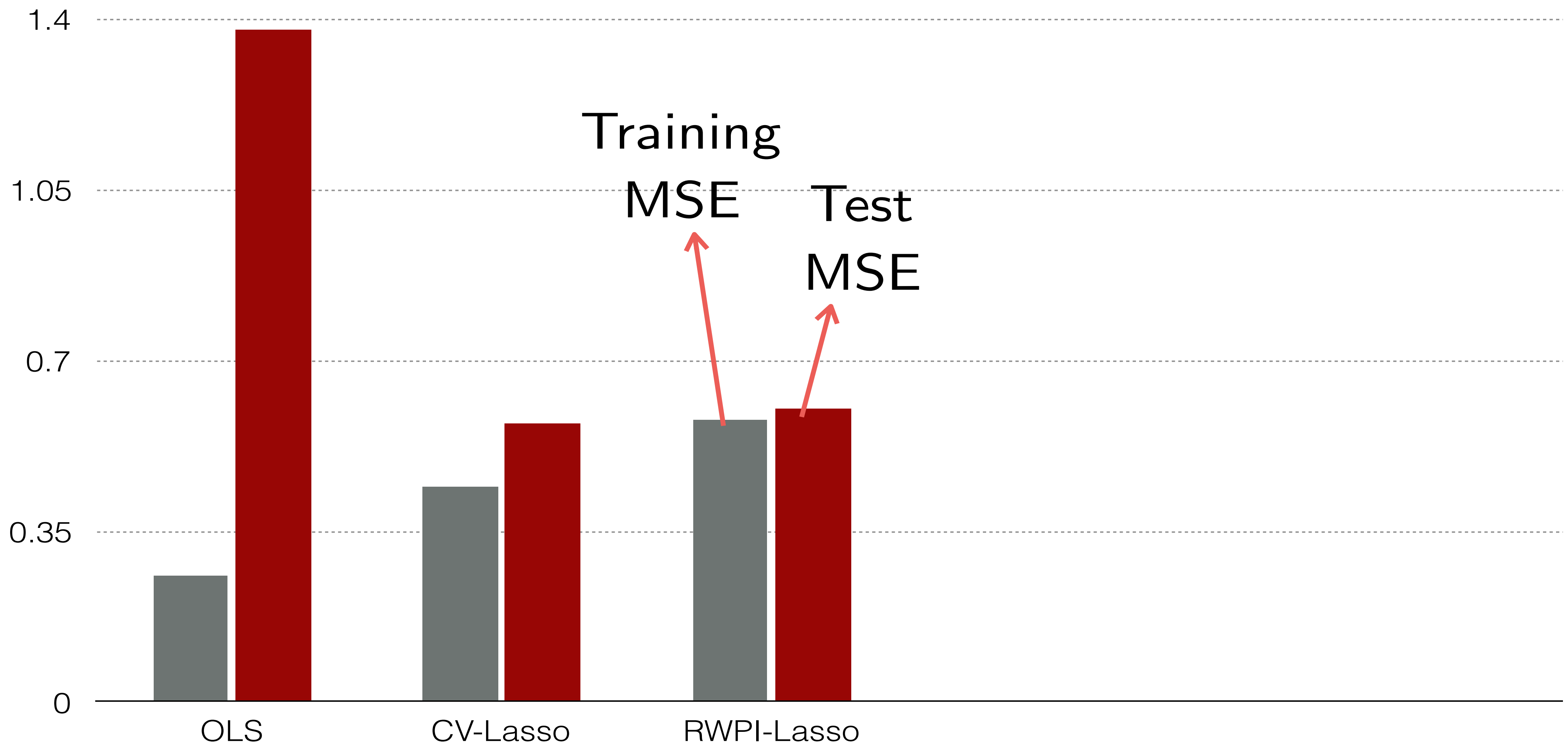
If $c(u,v) = \|u - v\|_q$,

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

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

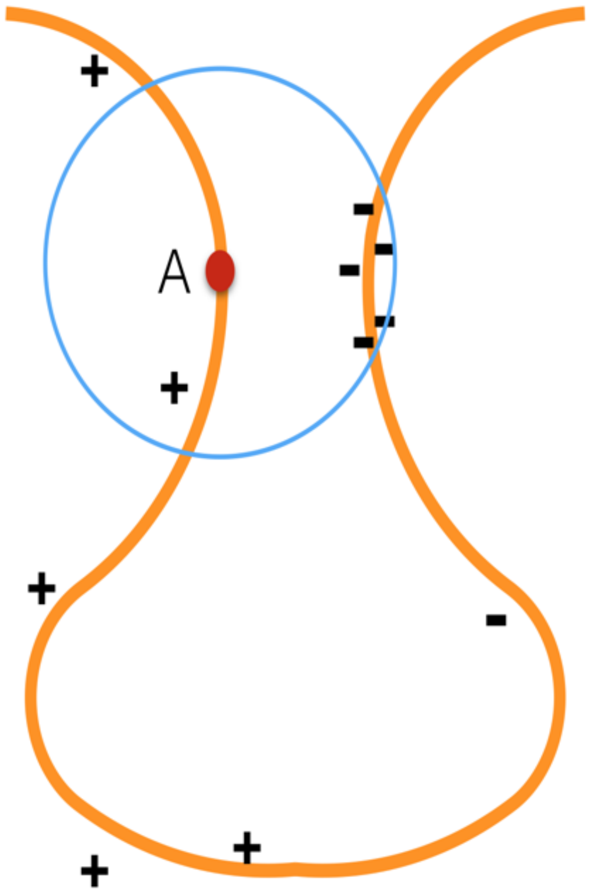
$$\delta \to \frac{\|Z\|_q}{\sqrt{n}}$$

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

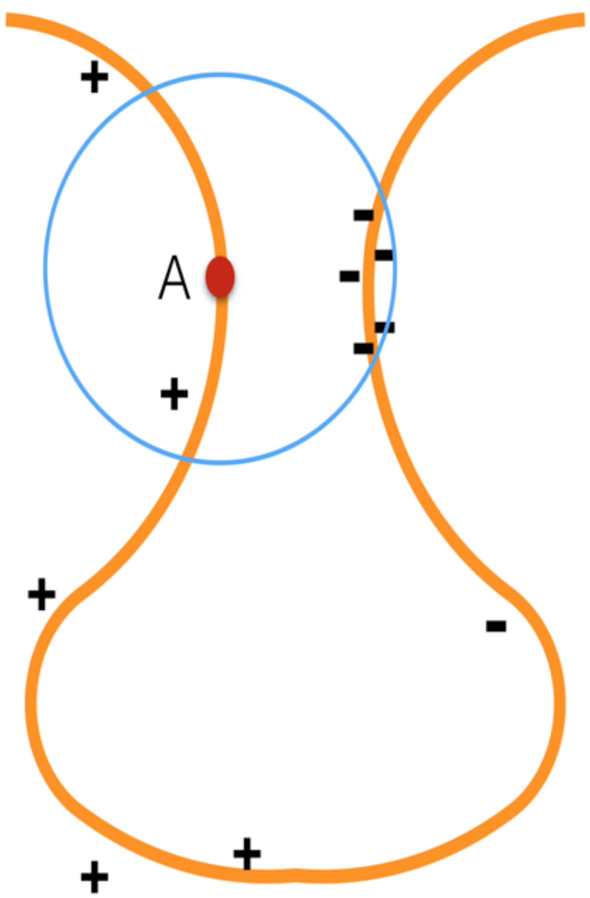


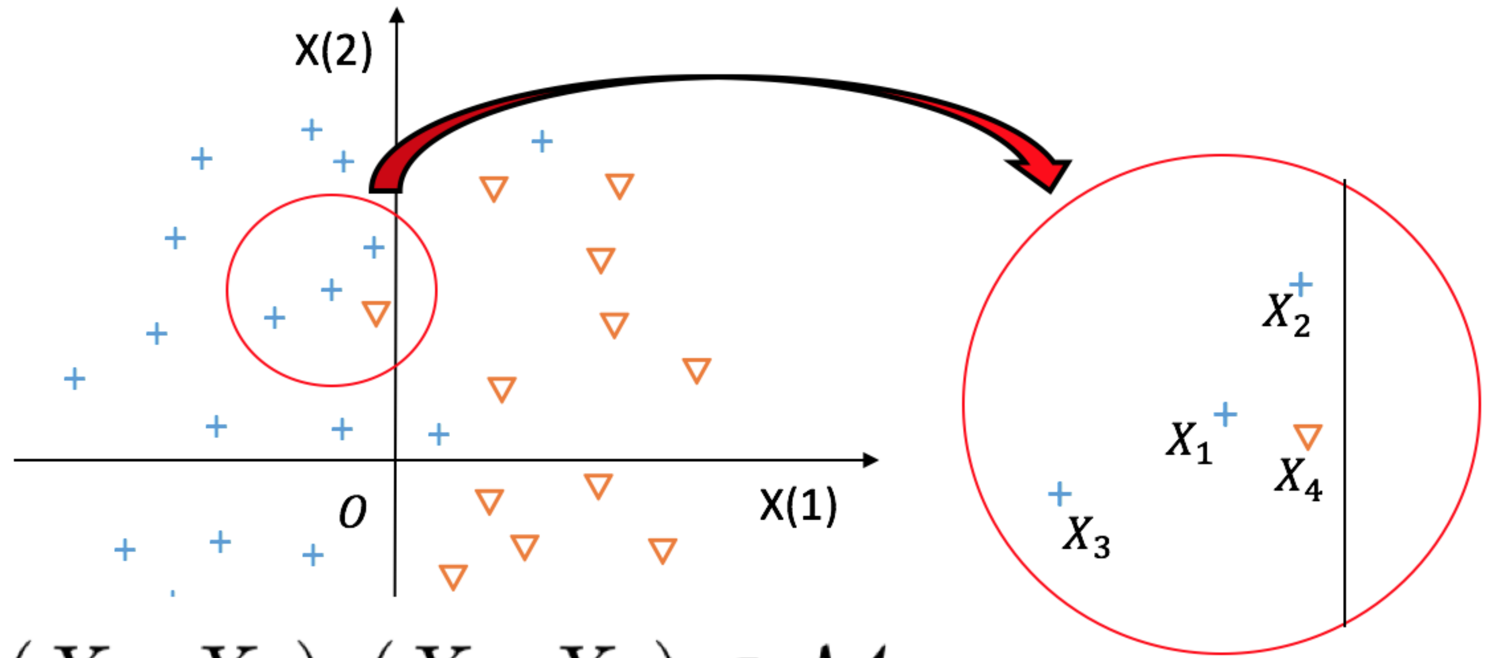
Limit result based radius choice vs cross-validation vs zero radius (OLS) in diabetic data set of 142 training samples with 64 predictors

Informing the geometry from data: Toy examples with classification



Informing the geometry from data: Toy examples with classification



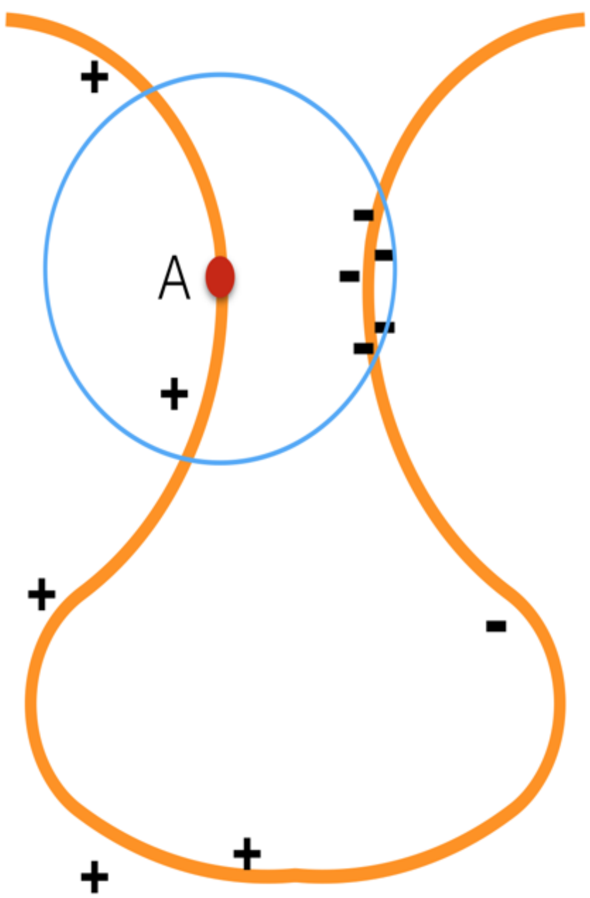


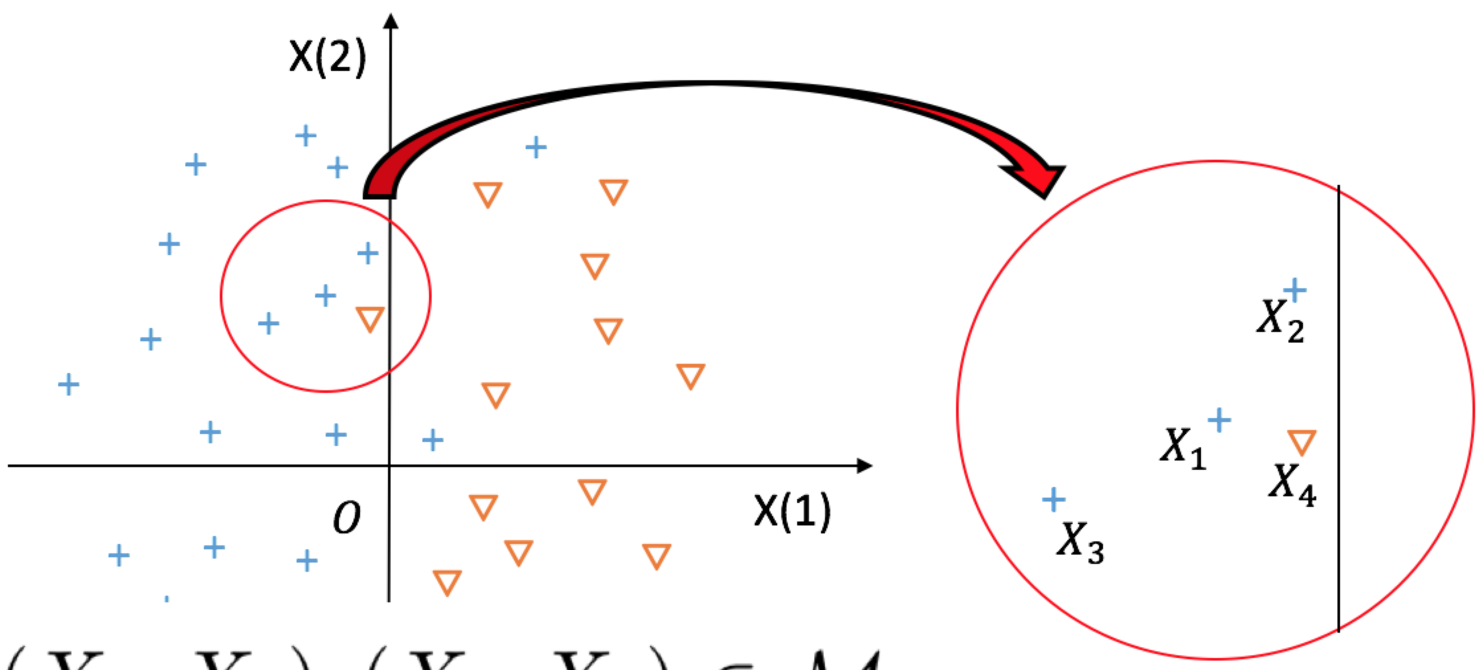
$$(X_1, X_2), (X_1, X_3) \in \mathcal{M}$$

$$(X_1, X_4) \in \mathcal{N}$$

$$\Lambda = \begin{bmatrix} 1.16 & 0 \\ 0 & 0.04 \end{bmatrix}$$

Informing the geometry from data: Toy examples with classification





$$(X_1, X_2), (X_1, X_3) \in \mathcal{M}$$

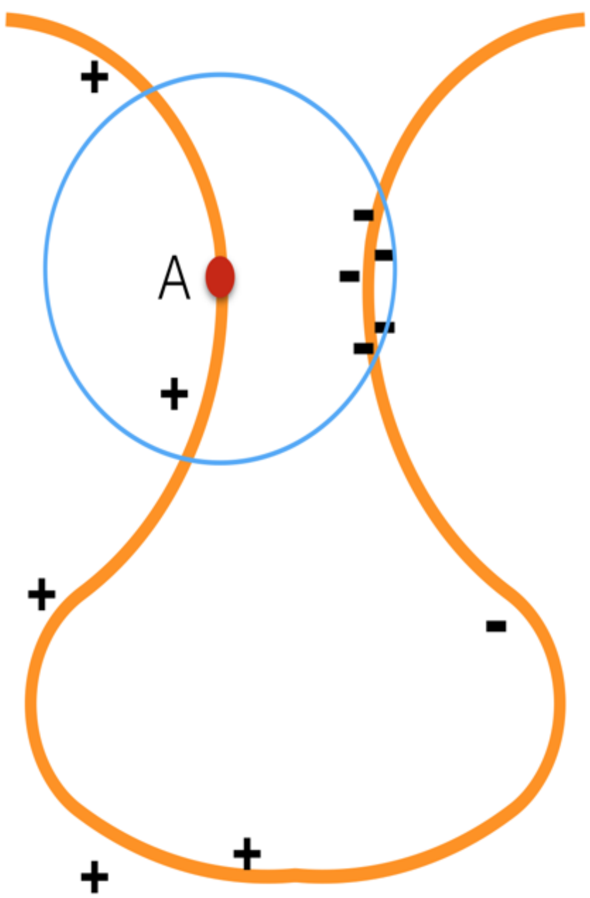
$$(X_1, X_4) \in \mathcal{N}$$

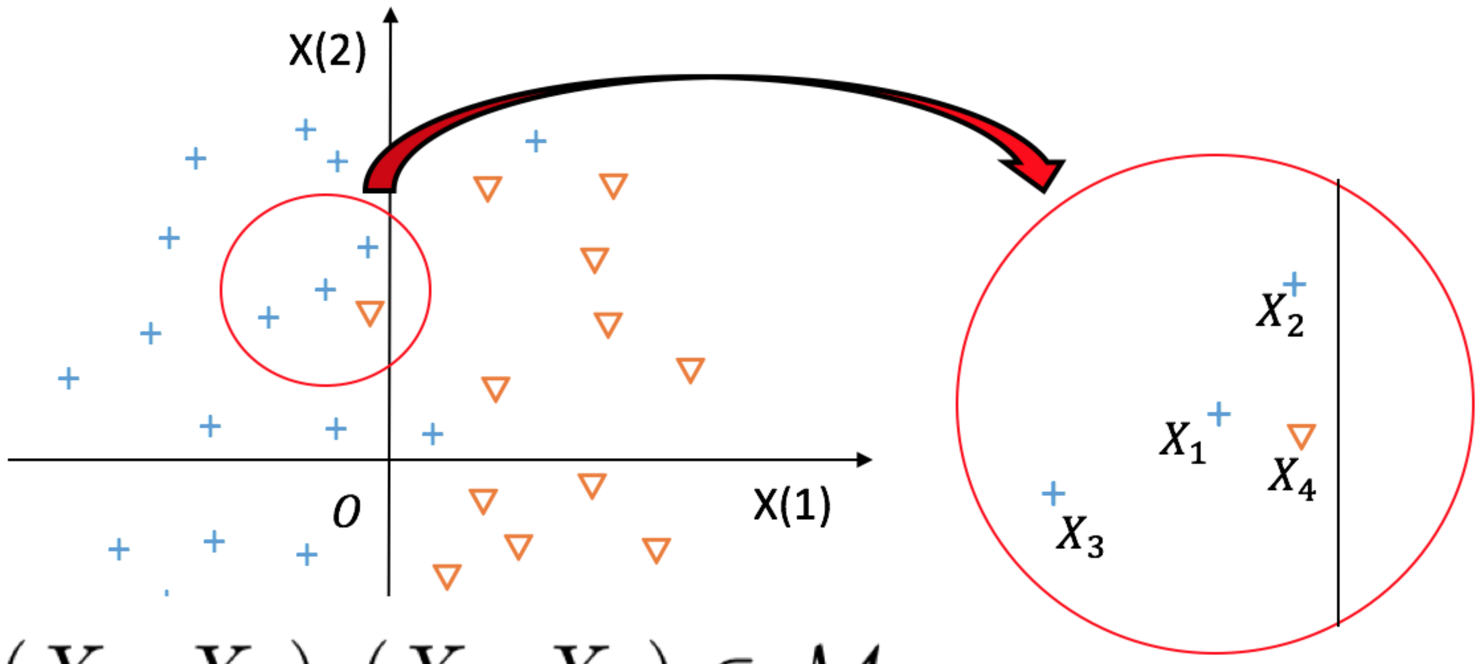
$$\Lambda = \begin{bmatrix} 1.16 & 0 \\ 0 & 0.04 \end{bmatrix}$$

$$\min_{\Lambda \in PSD} \sum_{(X_i, X_j) \in \mathcal{M}} d_\Lambda^2 (X_i, X_j)$$

$$s.t. \quad \sum_{(X_i, X_j) \in \mathcal{N}} d_\Lambda^2 (X_i, X_j) \geq \bar{\lambda}.$$

Informing the geometry from data: Toy examples with classification





$$(X_1, X_2), (X_1, X_3) \in \mathcal{M}$$

$$(X_1, X_4) \in \mathcal{N}$$

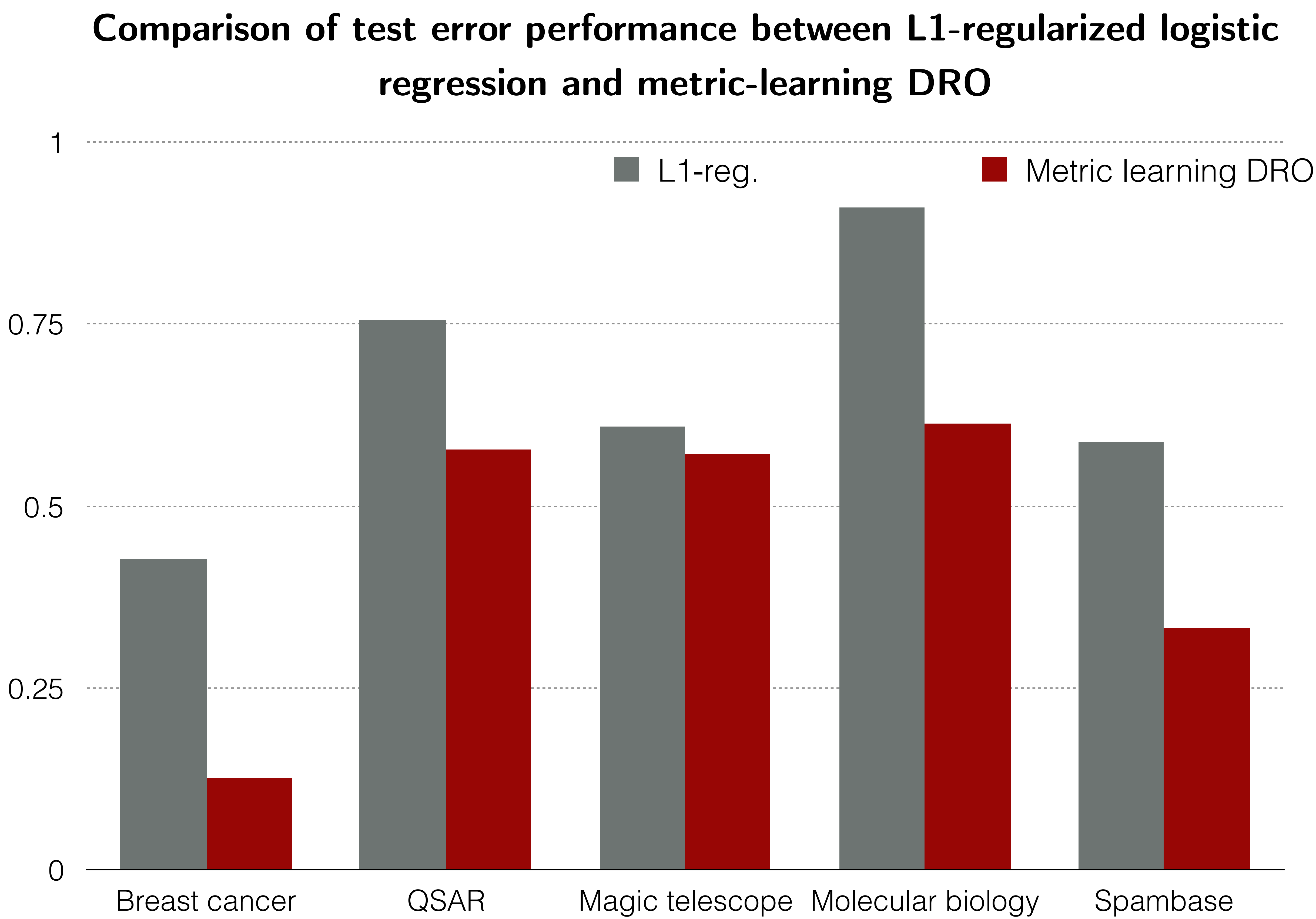
$$\Lambda = \begin{bmatrix} 1.16 & 0 \\ 0 & 0.04 \end{bmatrix}$$

$$\min_{\Lambda \in PSD} \sum_{(X_i, X_j) \in \mathcal{M}} d_\Lambda^2 (X_i, X_j)$$

$$s.t. \quad \sum_{(X_i, X_j) \in \mathcal{N}} d_\Lambda^2 (X_i, X_j) \geq \bar{\lambda}.$$

Take $c(x, y) = (x - y)^T \Lambda (x - y)$

$$\min_{\beta \in B} \sup_{P : D_c(P, P_n) \leq \delta} E_P \left[\ell(Y_i, \beta^T X_i) \right]$$



Comparison of test error performance between L1-regularized logistic regression and metric-learning DRO
1
0.75
0.5
0.25
0
L1-reg.
Metric learning DRO
Breast cancer
QSAR
Magic telescope
Molecular biology
Spambase

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

A flexible & attractive approach that allows

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

A flexible & attractive approach that allows

→ **to recast useful machine learning algorithms exactly as specific instances of (OT-DRO)**

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P\left[\ell(X;\beta)\right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

A flexible & attractive approach that allows

- to recast useful machine learning algorithms exactly as specific instances of (OT-DRO)
- a statistically principled approach towards selecting the radius of ambiguity region

$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P\left[\ell(X; \beta)\right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

A flexible & attractive approach that allows

- to recast useful machine learning algorithms exactly as specific instances of (OT-DRO)
- a statistically principled approach towards selecting the radius of ambiguity region
- scalable iterative schemes that are “at least as fast”, or “even faster” than the non-robust counterpart

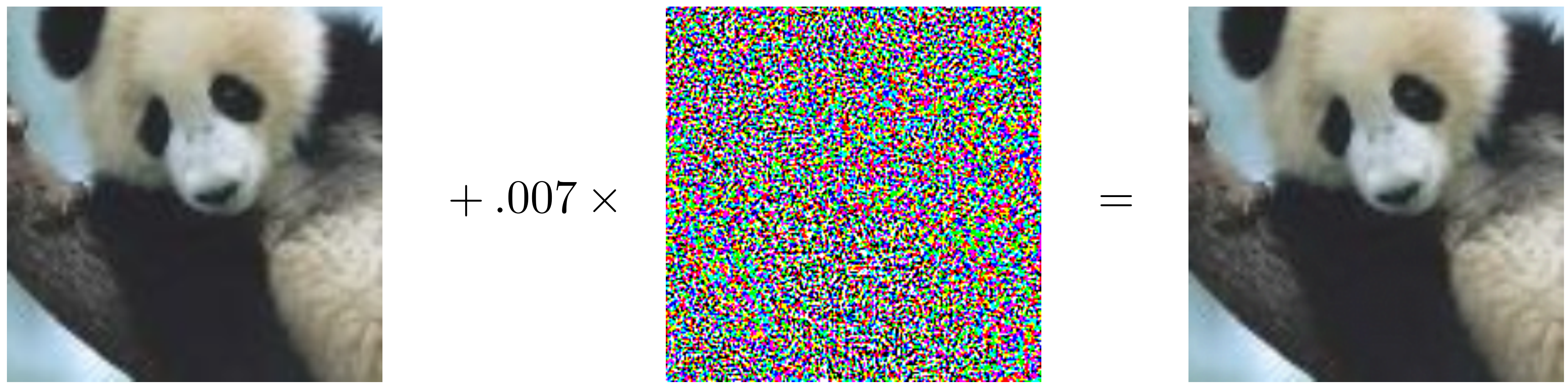
$$\inf_{\beta} \sup_{P: D(P, P_n) \leq \delta} E_P \left[\ell(X; \beta) \right] \qquad \text{(OT-DRO)}$$

Optimal mass transportation based DRO:

A flexible & attractive approach that allows

- to recast useful machine learning algorithms exactly as specific instances of (OT-DRO)
- a statistically principled approach towards selecting the radius of ambiguity region
- scalable iterative schemes that are “at least as fast”, or “even faster” than the non-robust counterpart
- the flexibility to inform the geometry of the ambiguity region from data and the improved performance it offers!

[Szegedy et al '15]



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

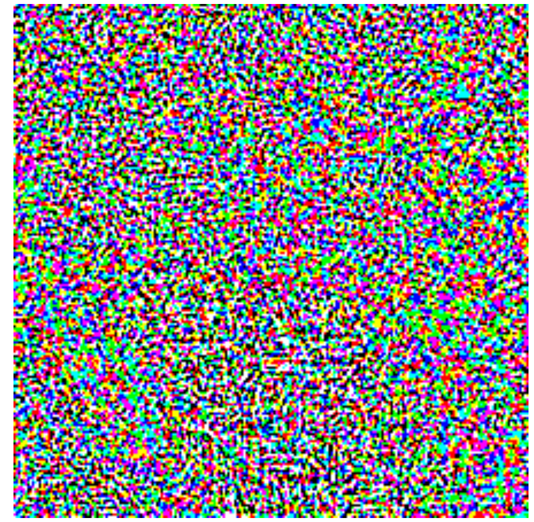
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[Szegedy et al '15]

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

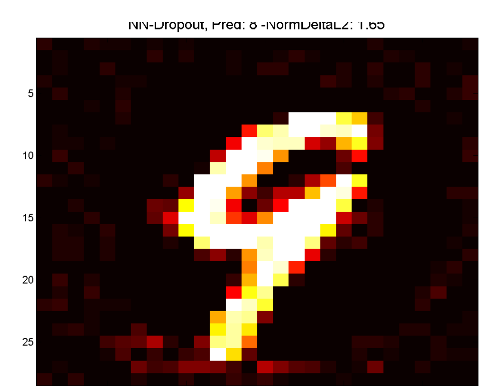
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

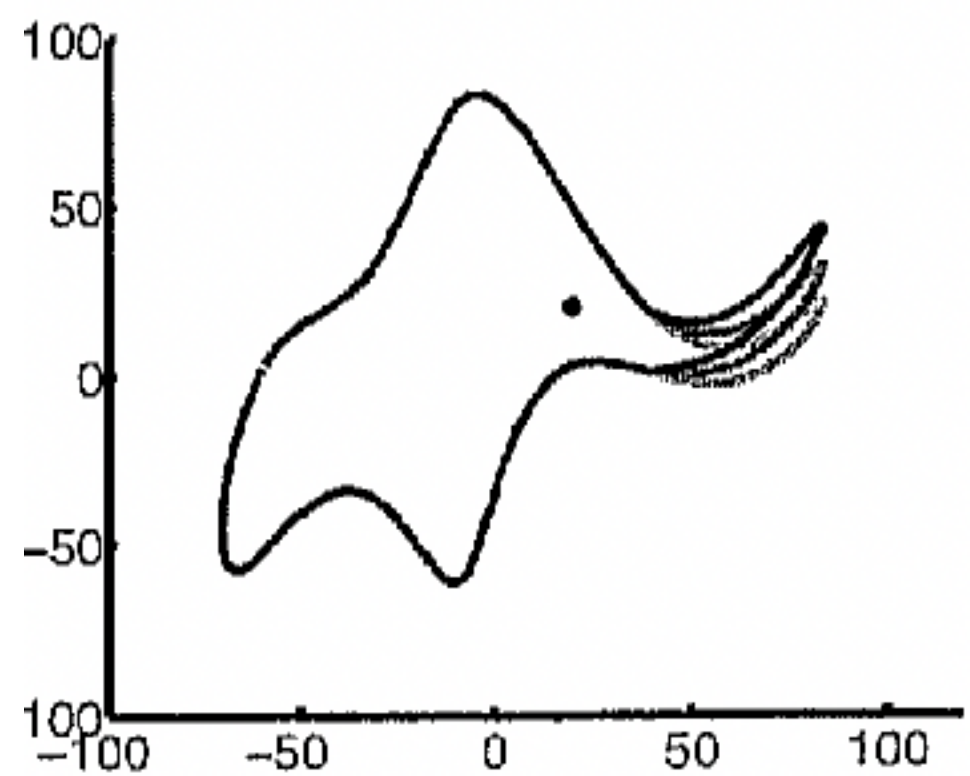
NN-WD, Pred:4, $\|\delta\|_2 = 1.7$



NN-DO, Pred:8, $\|\delta\|_2 = 1.7$

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

-von Neumann



Mayer et al '10

[Evtimov et al 2015]

[Evtimov et al 2015]

Some preprints

Quantifying distributional model risk via optimal transport
J Blanchet and K Murthy
2016 - https://arxiv.org/abs/1604.01446

Robust Wasserstein Profile Inference and its applications to Machine learning
J Blanchet, Y Kang and K Murthy
2016 - https://arxiv.org/abs/1610.05627

Data-driven optimal cost selection for Distributionally Robust Optimization
J Blanchet, Y Kang, F Zhang and K Murthy
2017 - https://arxiv.org/pdf/1705.07152.pdf

Stochastic gradient descent for Optimal transport DRO
J Blanchet, K Murthy and F Zhang
(To be available soon)