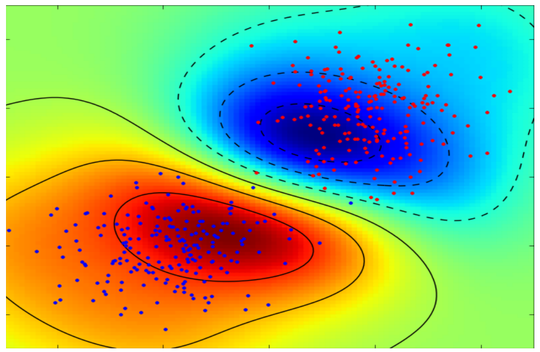
Regularization via Optimal Transport

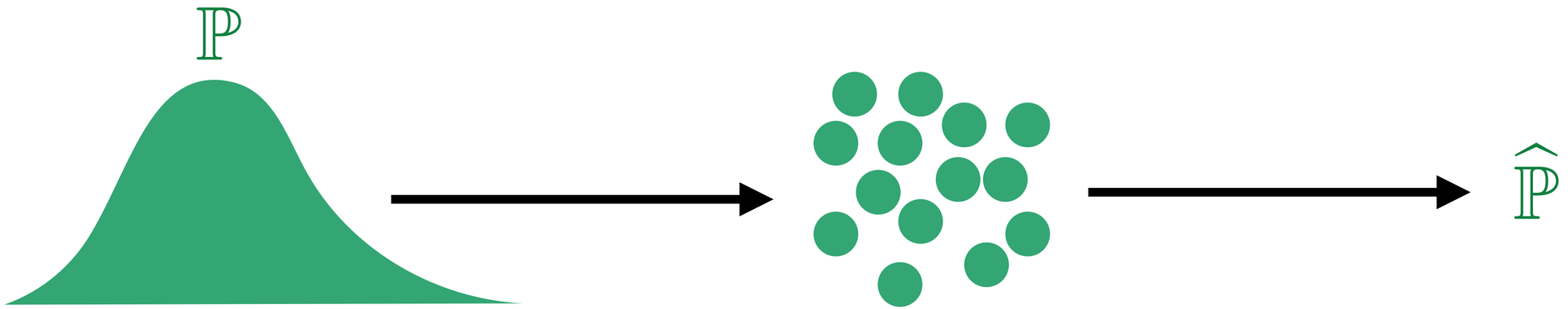
Soroosh Shafieezadeh Abadeh, Tepper School of Business, CMU

Stochastic Programming

$$\inf_{\theta \in \Theta} \; \mathbb{E}_{\mathbb{P}} \left[\ell(\theta, \xi) \right]$$



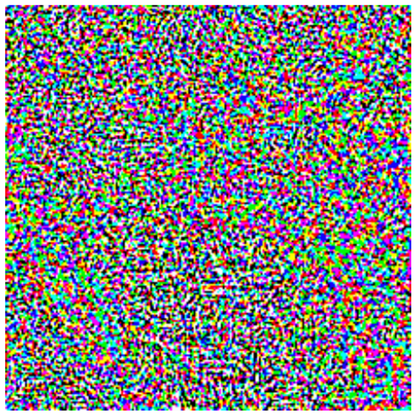
Failure Examples: Overfitting



$$\inf_{\theta \in \Theta} \; \mathbb{E}_{\widehat{\mathbb{P}}} \left[\ell(\theta, \xi) \right]$$

Failure Examples: Adversarial Attack [GSS15]

$+ .007 \times$



$=$

Panda
57.7% Confidence

Gibbon
99.3% Confidence

Failure Examples: Fake Data

TECH | PERSONAL TECH | PERSONAL TECHNOLOGY: NICOLE NGUYEN

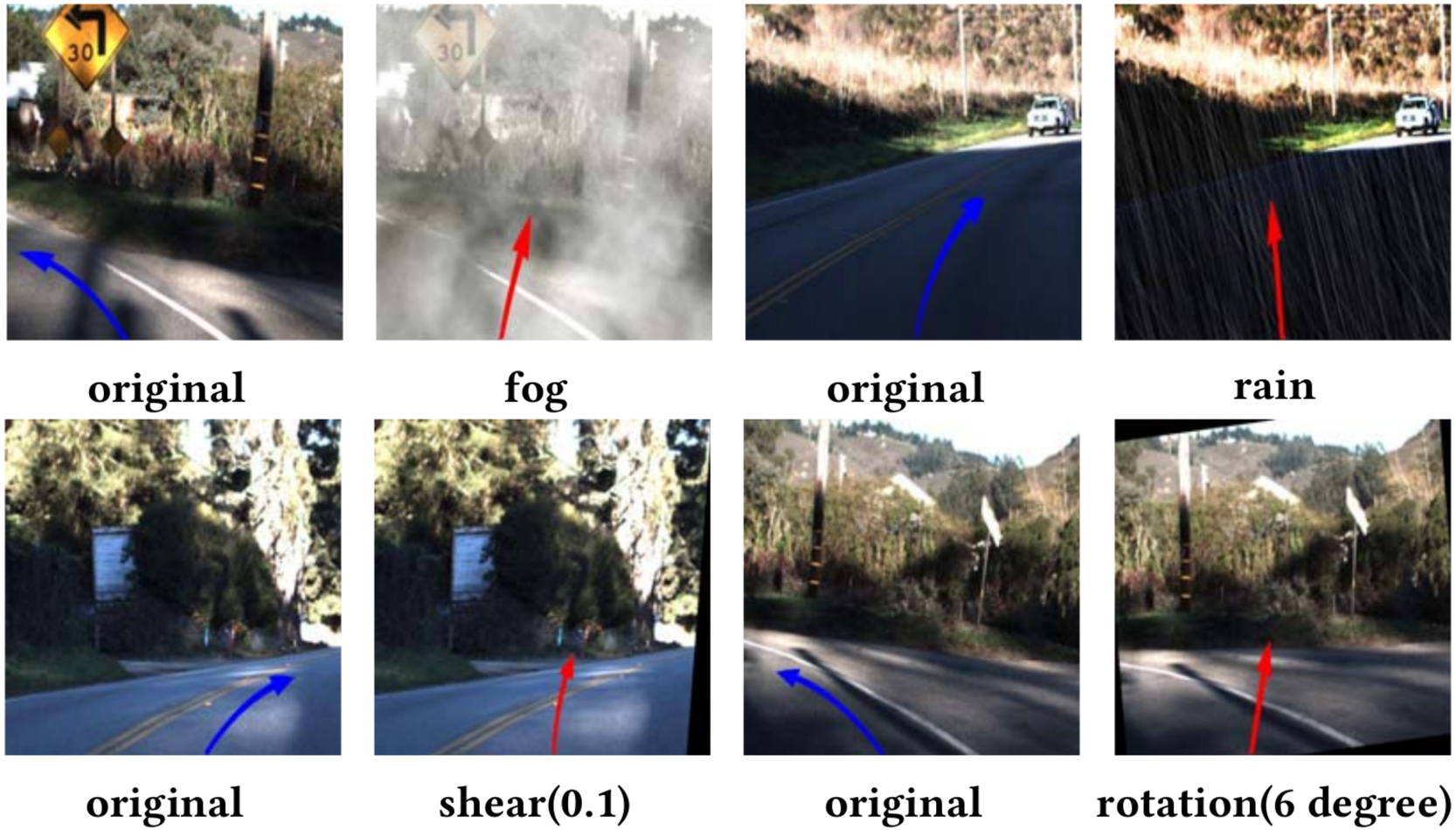
Fake Reviews and Inflated Ratings Are Still a Problem for Amazon

Sellers are taking advantage of the online-shopping frenzy, using old and new methods to boost ratings on products

By Nicole Nguyen
June 13, 2021 8:28 am ET

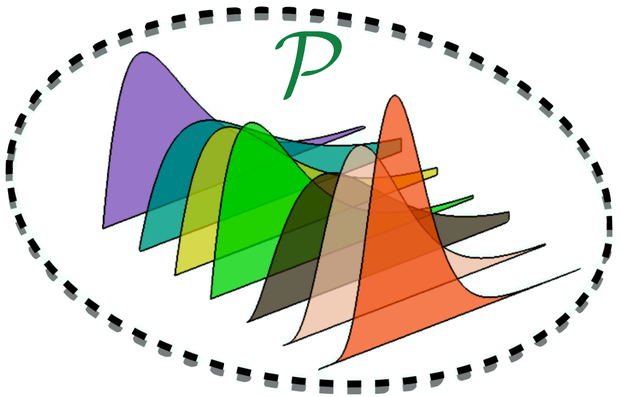
A charging brick recently caught my eye on Amazon. AMZN -2.96% ▼ It was a RAVPower-branded two-port fast charger, and it had five stars with over 9,800 ratings. The score seemed suspect but Amazon itself

Failure Examples: Domain Change [TPJR18]

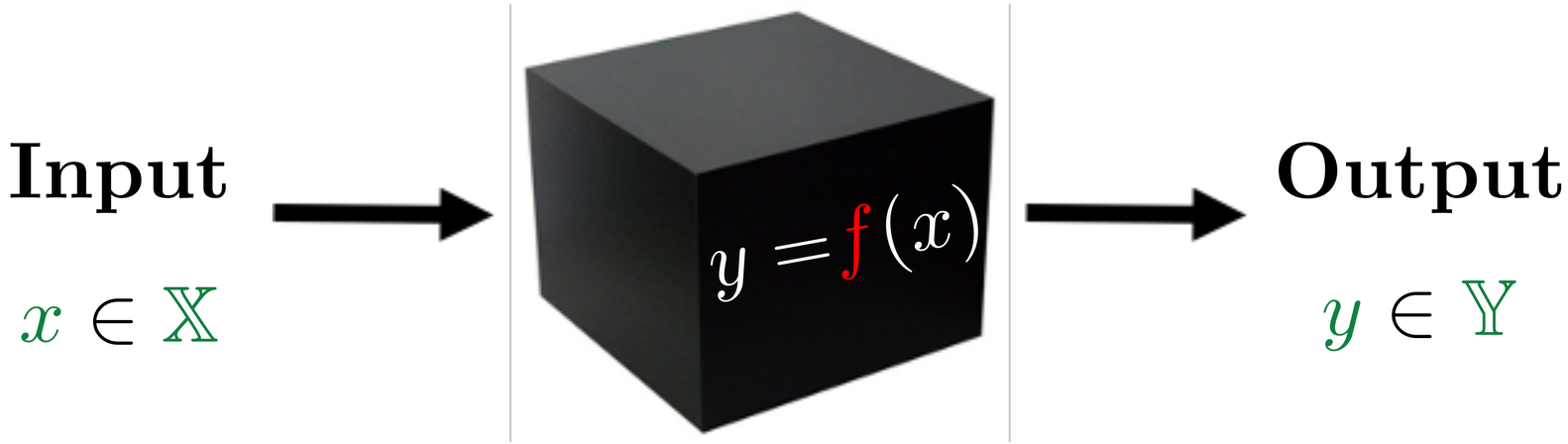


Distributionally Robust Optimization

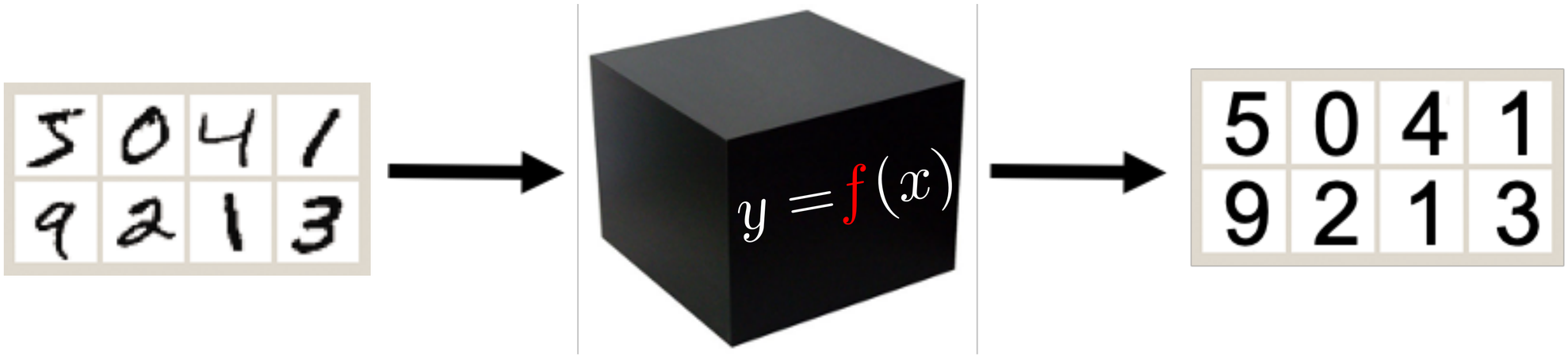
$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right]$$



Supervised Learning



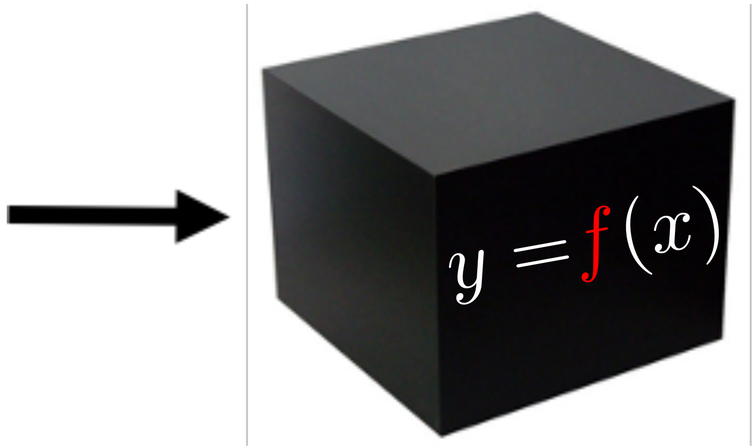
Supervised Learning



Supervised Learning

Input

$x \in \mathbb{X}$

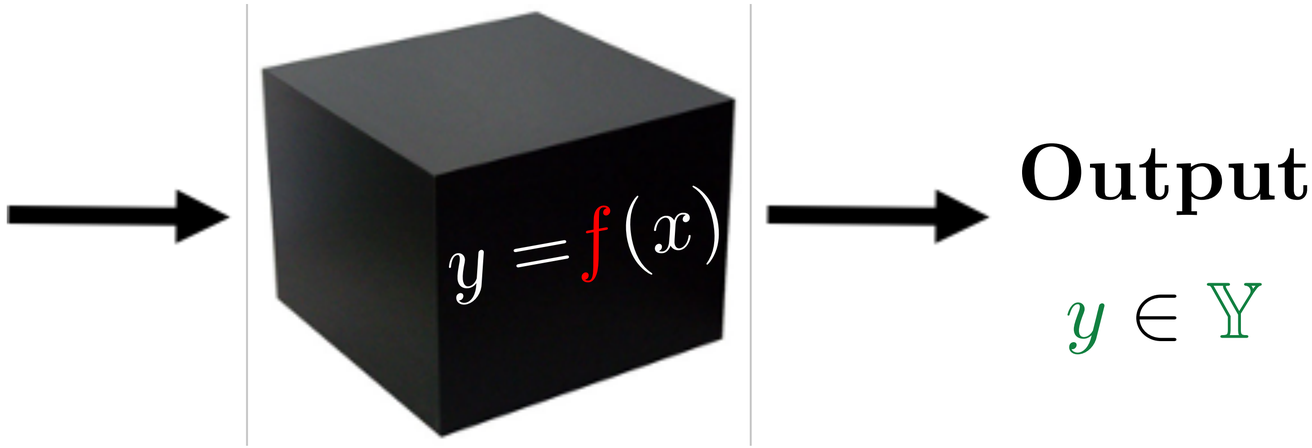


Output

$y \in \mathbb{Y}$

Training data: $\widehat{\Xi}_N = (\widehat{x}_1, \widehat{y}_1), \ldots (\widehat{x}_N, \widehat{y}_N)$

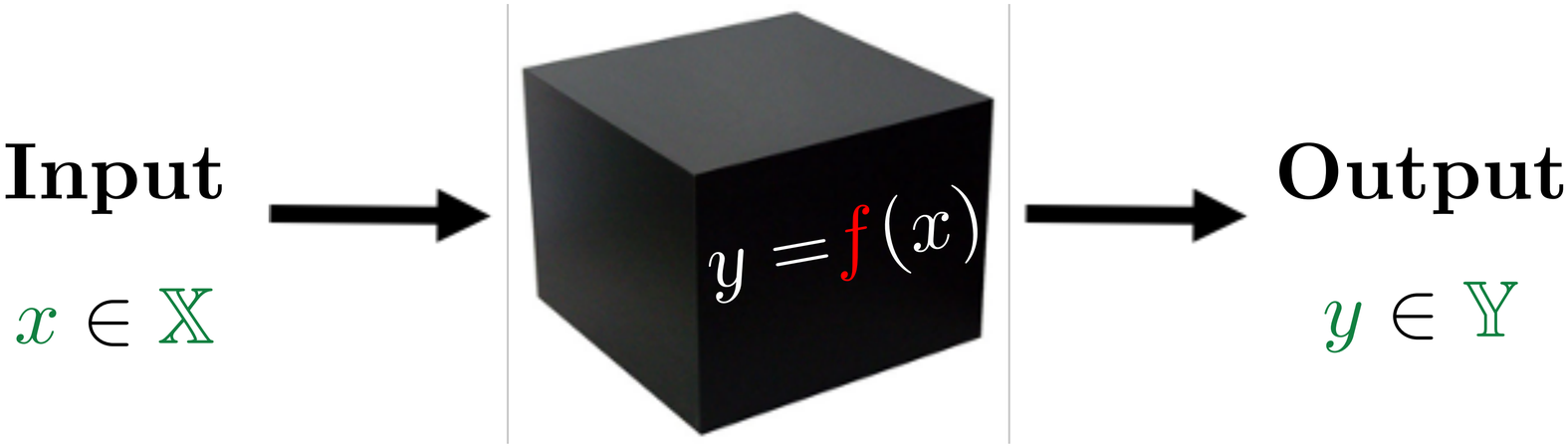
Supervised Learning



Training data: $\widehat{\Xi}_N = (\widehat{x}_1, \widehat{y}_1), \ldots (\widehat{x}_N, \widehat{y}_N)$

Hypothesis space: $\mathbb{H} \subseteq \{h \in \mathbb{R}^{\mathbb{X}}\}$

Supervised Learning

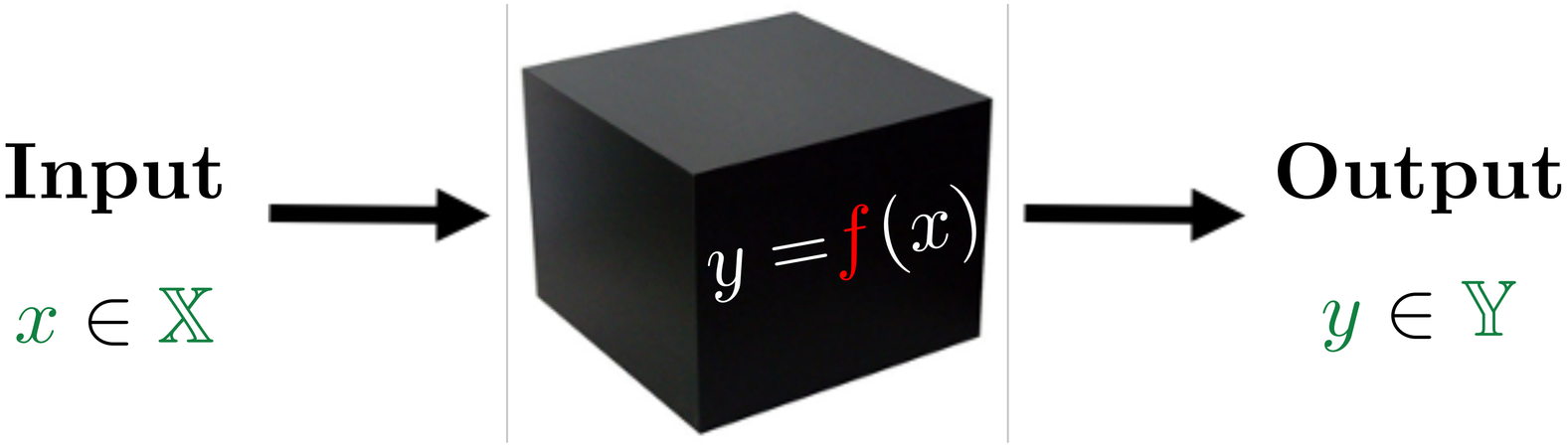


Training data: $\widehat{\Xi}_N = (\widehat{x}_1, \widehat{y}_1), \ldots (\widehat{x}_N, \widehat{y}_N)$

Hypothesis space: $\mathbb{H} \subseteq \{h \in \mathbb{R}^{\mathbb{X}}\}$

Target function: $f(x) \approx h(x)$

Supervised Learning



Training data: $\widehat{\Xi}_N = (\widehat{x}_1, \widehat{y}_1), \ldots (\widehat{x}_N, \widehat{y}_N)$

Hypothesis space: $\mathbb{H} \subseteq \{h \in \mathbb{R}^{\mathbb{X}}\}$

Target function: $f(x) \approx h(x)$

Learning algorithm: $\inf_{h \in \mathbb{H}} \ell(h, \widehat{\Xi}_N)$

Regression Models

Target function: $f(x) = h(x)$

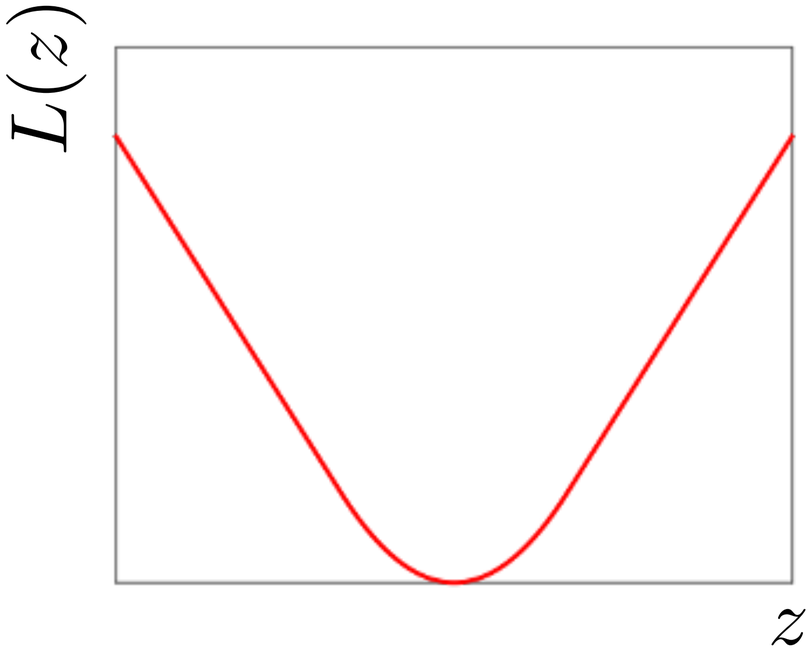
Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i) - \widehat{y}_i)$

Regression Models

Target function: $f(x) = h(x)$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i) - \widehat{y}_i)$

Robust Regression



Huber Loss:

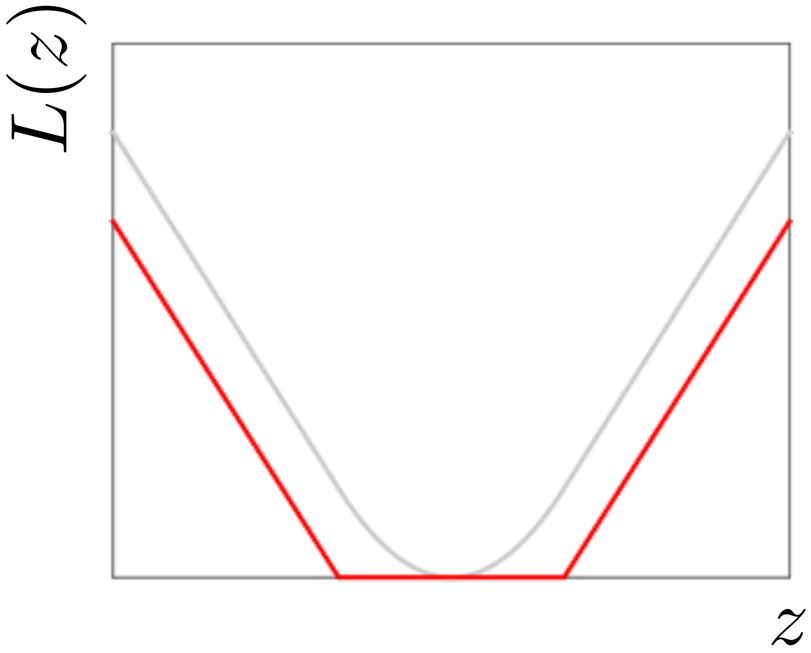
$$L(z) = \begin{cases} \frac{1}{2} z^2 & \text{if } |z| \leq \delta \\ \delta \left(|z| - \frac{1}{2} \delta \right) & \text{else} \end{cases}$$

Regression Models

Target function: $f(x) = h(x)$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i) - \widehat{y}_i)$

Support Vector Regression



ε-insensitive Loss:

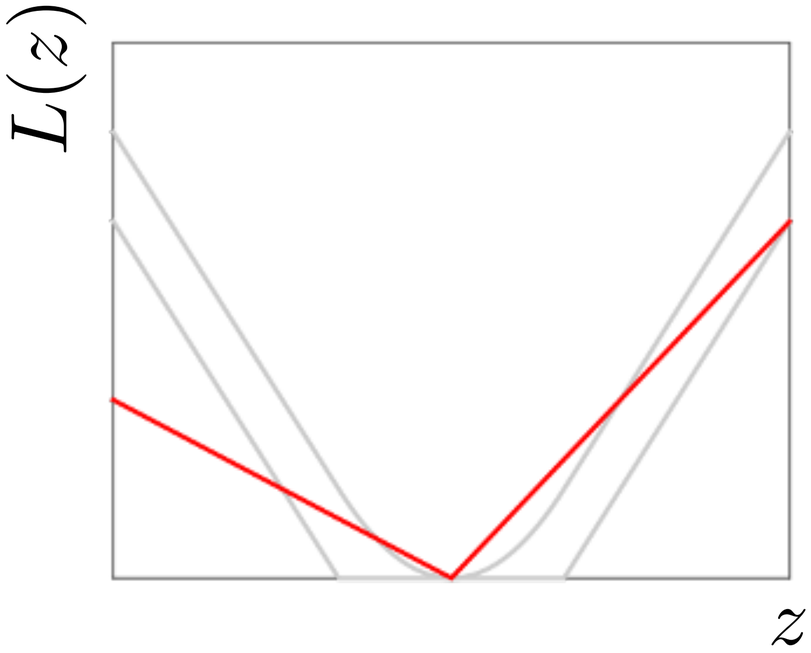
$$L(z) = \max\{0, |z| - \varepsilon\}$$

Regression Models

Target function: $f(x) = h(x)$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i) - \widehat{y}_i)$

Quantile Regression



Pinball Loss:

$$L(z) = \max\{-\tau z, (1-\tau)z\}$$

Classification Models

Target function: $f(x) = \operatorname{sgn}(h(x))$

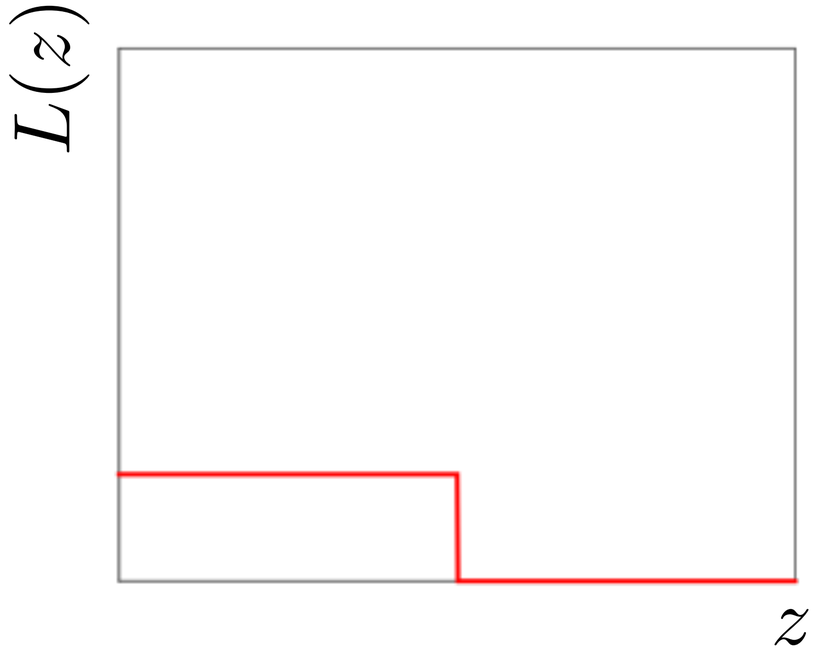
Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} L(\widehat{y}_i h(\widehat{x}_i))$

Classification Models

Target function: $f(x) = \operatorname{sgn}(h(x))$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} L(\widehat{y}_i h(\widehat{x}_i))$

Ideal Classification



0-1 Loss:

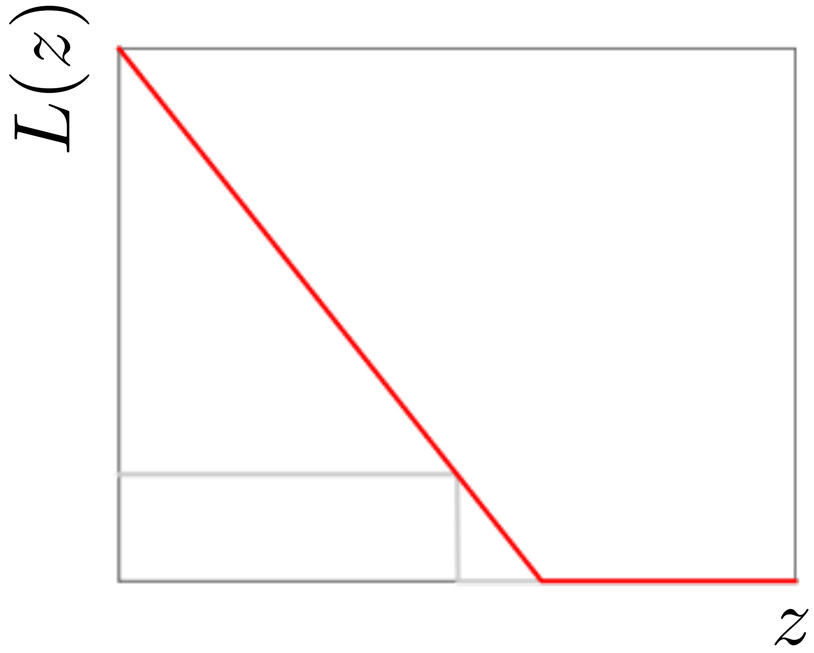
$$L(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 0 & \text{else} \end{cases}$$

Classification Models

Target function: $f(x) = \text{sgn}(h(x))$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(\widehat{y}_i h(\widehat{x}_i))$

Support Vector Machine



Hinge Loss:

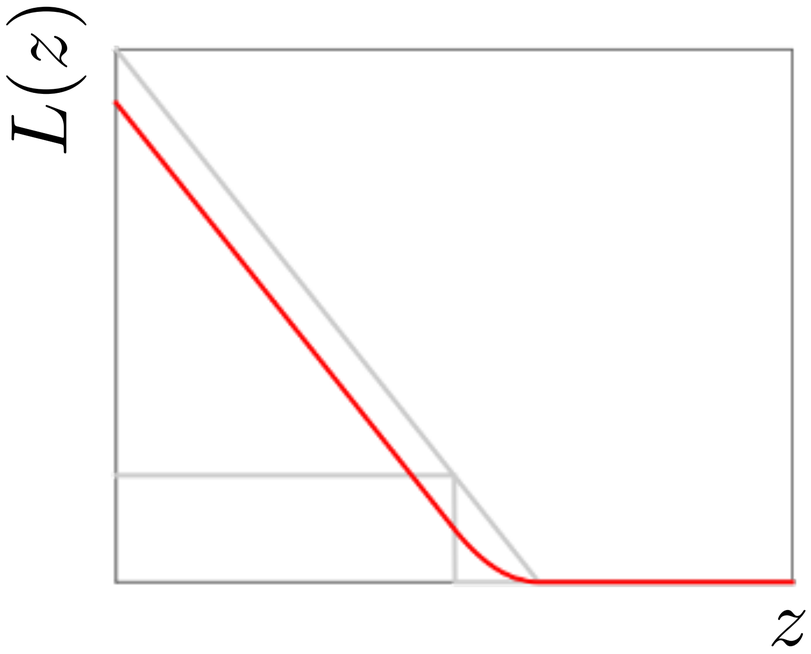
$$L(z) = \max\{0, 1 - z\}$$

Classification Models

Target function: $f(x) = \operatorname{sgn}(h(x))$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \dfrac{1}{N} \sum_{i=1}^{N} L(\widehat{y}_i h(\widehat{x}_i))$

Support Vector Machine II



Smooth Hinge Loss:

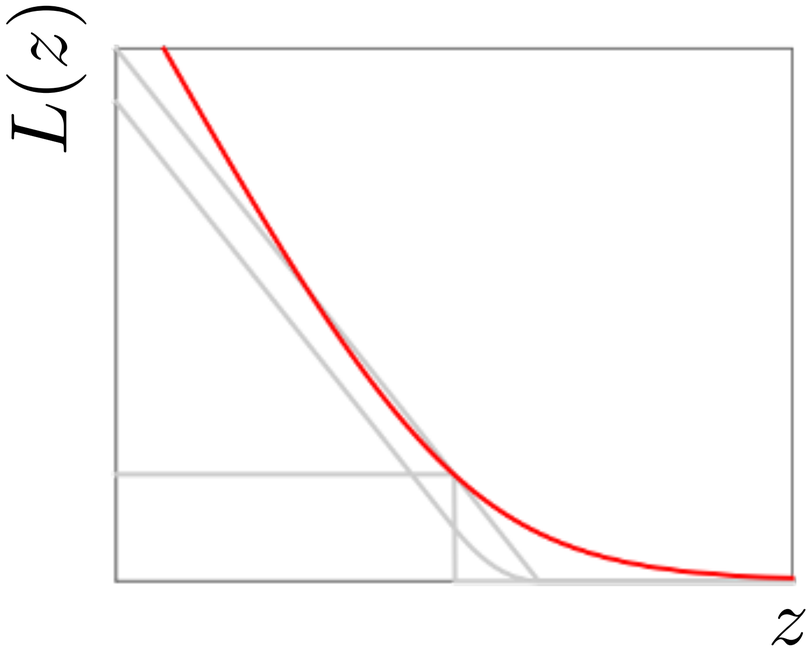
$$L(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1-z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{else} \end{cases}$$

Classification Models

Target function: $f(x) = \text{sgn}(h(x))$

Empirical risk minimization: $\ell(h, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^{N} L(\widehat{y}_i h(\widehat{x}_i))$

Logistic Regression



Logloss:

$$L(z) = \log(1 + \exp(-z))$$

Performance of ERM

$$h_{\mathrm{ERM}} = \underset{h \in \mathbb{H}}{\operatorname{argmin}} \ \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i), \widehat{y}_i)$$

In-Sample Performance:

$$\mathbb{E}_{\widehat{\mathbb{P}}_N} \left[L(h(\widehat{x}_i), \widehat{y}_i) \right]$$

Performance of ERM

$$h_{\mathrm{ERM}} = \underset{h \in \mathbb{H}}{\operatorname{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i), \widehat{y}_i)$$

In-Sample Performance:

$$\mathbb{E}_{\widehat{\mathbb{P}}_N} \left[L(h(\widehat{x}_i), \widehat{y}_i) \right]$$

Out-of-Sample Performance:

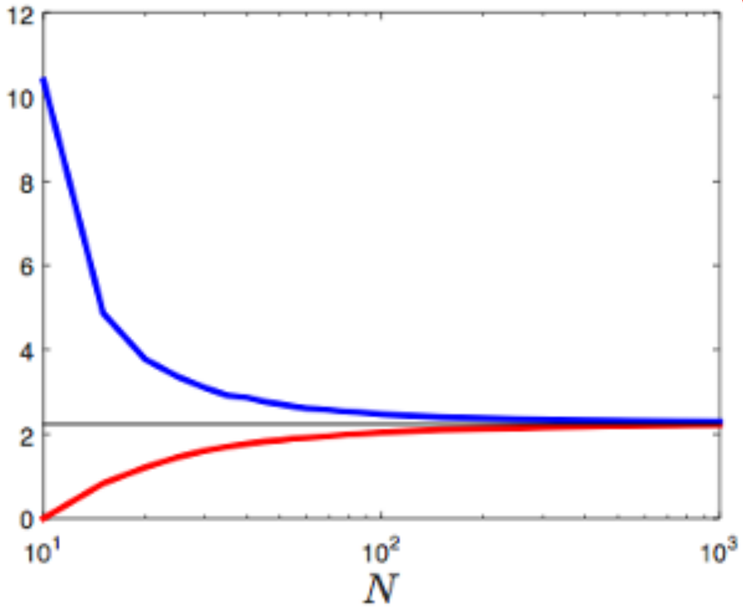
$$\mathbb{E}_{\mathbb{P}} \left[L(h(\widehat{x}_i), \widehat{y}_i) \right]$$

Performance of ERM

$$h_{\mathrm{ERM}} = \underset{h \in \mathbb{H}}{\operatorname{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i), \widehat{y}_i)$$

In-Sample Performance:

$$\mathbb{E}_{\widehat{\mathbb{P}}_N} \left[L(h(\widehat{x}_i), \widehat{y}_i) \right]$$



Out-of-Sample Performance:

$$\mathbb{E}_{\mathbb{P}} \left[L(h(\widehat{x}_i), \widehat{y}_i) \right]$$

Regularized ERM

$$h_{\mathrm{REG}} = \underset{h \in \mathbb{H}}{\operatorname{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i), \widehat{y}_i) + \varepsilon \, \Omega(h)$$

Regularization coefficient

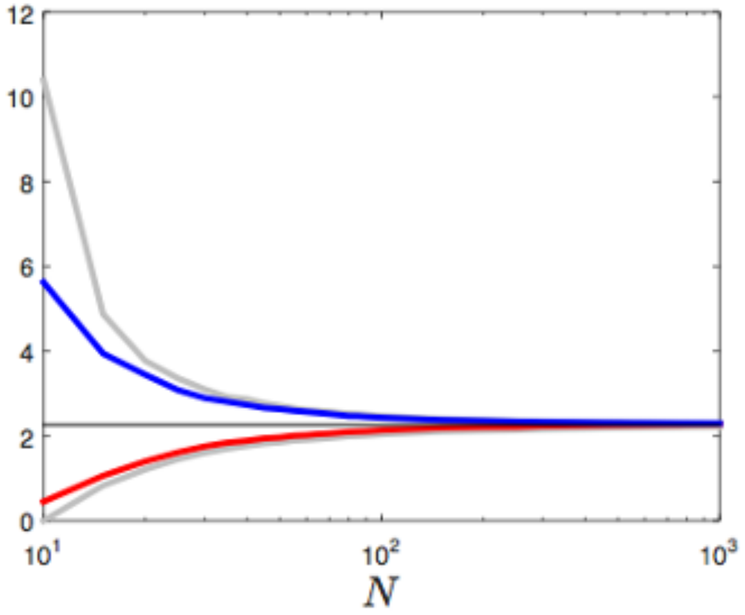
Regularization function

Regularized ERM

$$h_{\mathrm{REG}} = \underset{h \in \mathbb{H}}{\operatorname{argmin}} \; \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{x}_i), \widehat{y}_i) + \varepsilon \, \Omega(h)$$

Regularization coefficient

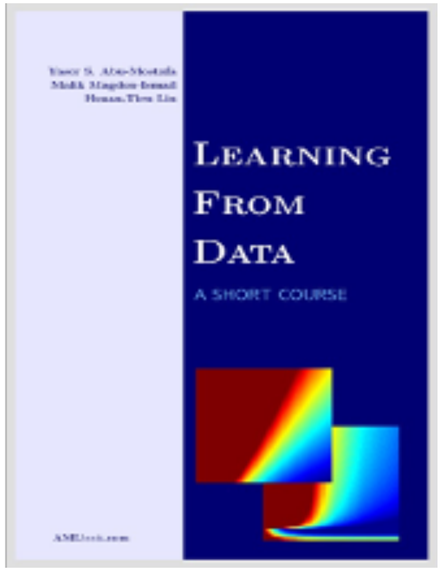
Regularization function



Regularized ERM

"Most of the **regularization methods** used successfully in practice **are heuristic methods**."

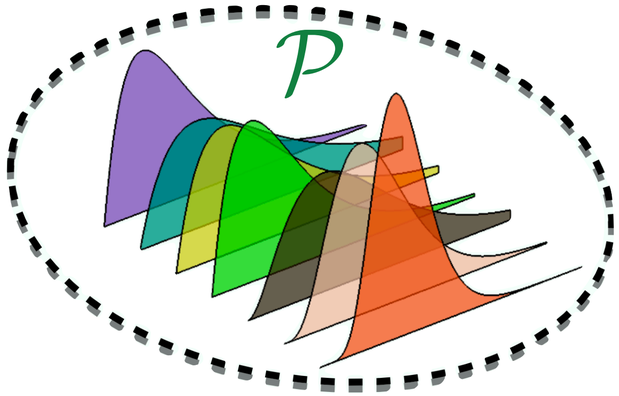
Abu-Mostafa *et al.*, 2012.



Regularization via Optimal Transport

$$\inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[L(h(x), y) \right]$$

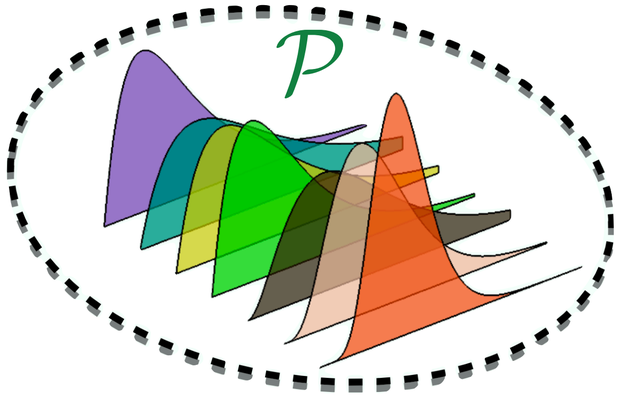
$$\mathbb{H} = \left\{ h \in \mathbb{R}^{\mathbb{X}} : \exists \theta \in \Theta \text{ s.t. } h(x) = \theta^{\top} x \right\}$$



Regularization via Optimal Transport

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[L(\theta^\top x, y) \right]$$

$$\mathbb{H} = \left\{ h \in \mathbb{R}^{\mathbb{X}} : \exists \theta \in \Theta \text{ s.t. } h(x) = \theta^\top x \right\}$$



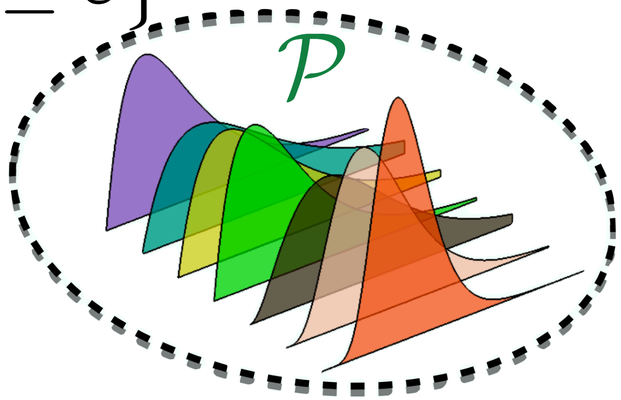
Regularization via Optimal Transport

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[L(\theta^\top x, y) \right]$$

$$\mathbb{H} = \left\{ h \in \mathbb{R}^{\mathbb{X}} : \exists \theta \in \Theta \text{ s.t. } h(x) = \theta^\top x \right\}$$

$$\mathcal{P} = \left\{ \mathbb{Q} \in \mathcal{M}(\mathbb{R}^n \times \mathbb{Y}) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$$

$$\widehat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\widehat{x}_i, \widehat{y}_i)}$$



The Real Story Behind the Success

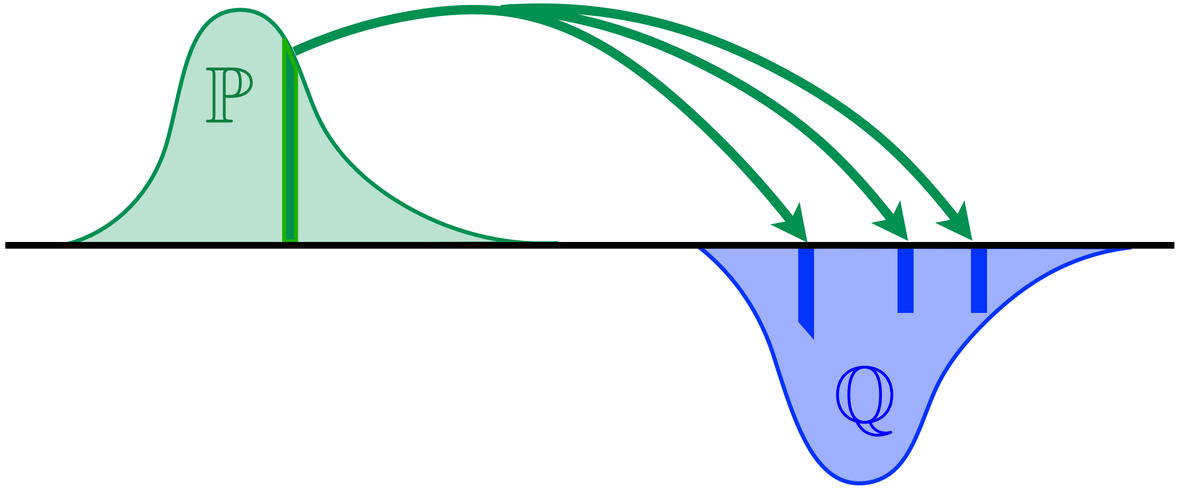
Regularization

[SMK15, GCK20, CP18, BMZ18, BKM19, SKM19]

Statistical Guarantees

[SMK15, MK18, BKM19, SKM19, Gao20, BMN21]

What is Optimal Transport?



$$W_c(\mathbb{Q},\mathbb{P}) = \begin{cases} \inf_{\pi\in\mathcal{M}(\Xi,\Xi)} & \mathbb{E}_\pi\left[c(\xi,\xi')\right] \\ \text{s.t.} & \pi \in \Pi(\mathbb{Q},\mathbb{P}) \end{cases}$$

THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES **1941**

By Frank L. Hitchcock

1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from city to city.

OPTIMUM UTILIZATION OF THE TRANSPORTATION SYSTEM* **1949**

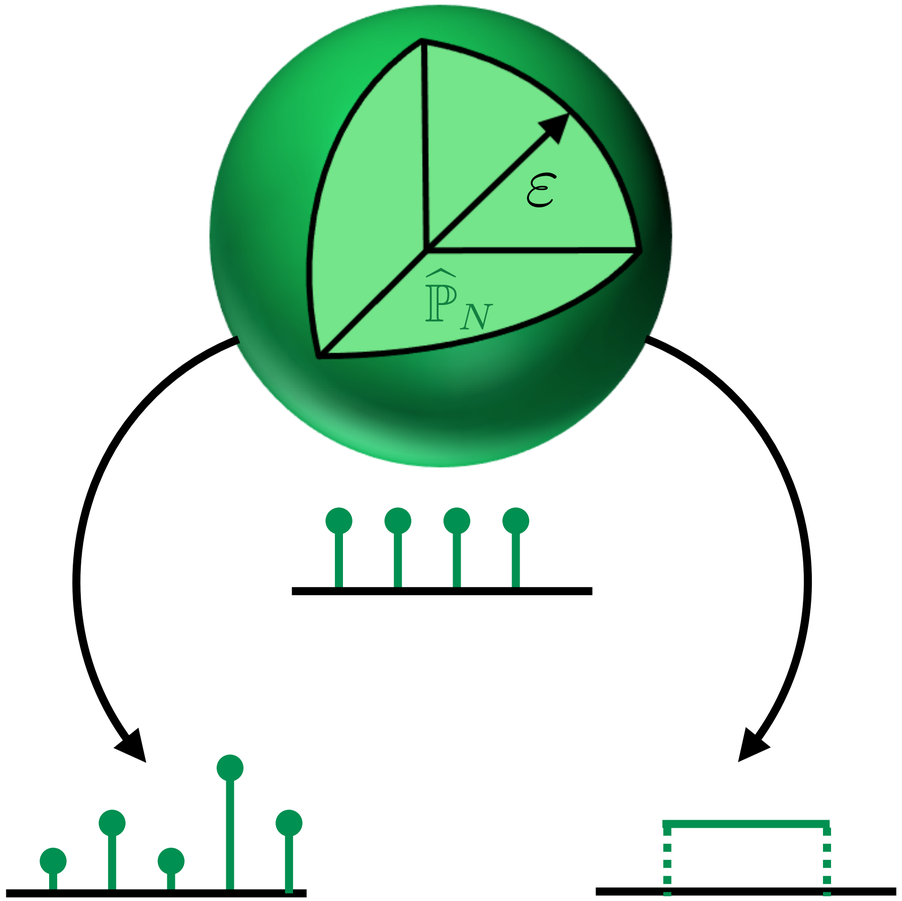
by Tjalling C. Koopmans

Professor of Economics, The University of Chicago, and Research Associate, Cowles Commission for Research in Economics

The purpose of this paper is to give an application of the theory of optimum allocation of resources to one particular industry. I shall, therefore, not speak on that theory in general. I shall use one of its

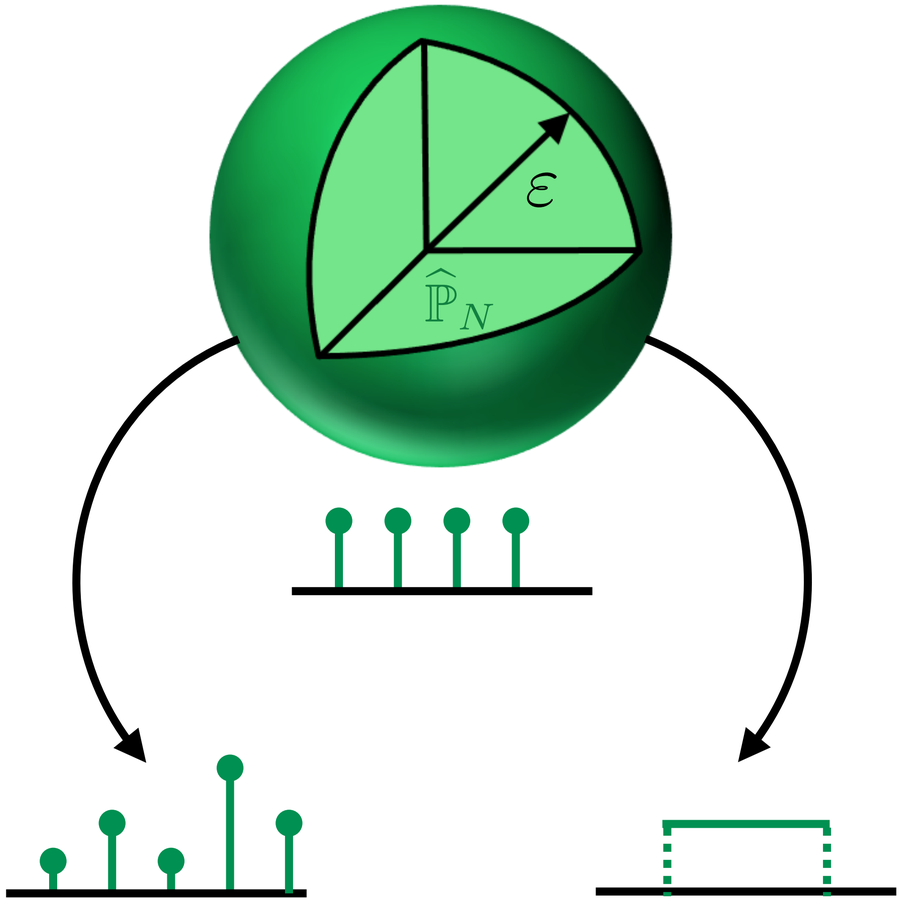
Optimal Transport Ambiguity Set

$$\mathcal{P} = \left\{ \mathbb{Q} \in \mathcal{M}(\mathbb{R}^n \times \mathbb{Y}) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$$



Optimal Transport Ambiguity Set

$$\mathcal{P} = \left\{ \mathbb{Q} \in \mathcal{M}(\mathbb{R}^n \times \mathbb{Y}) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$$



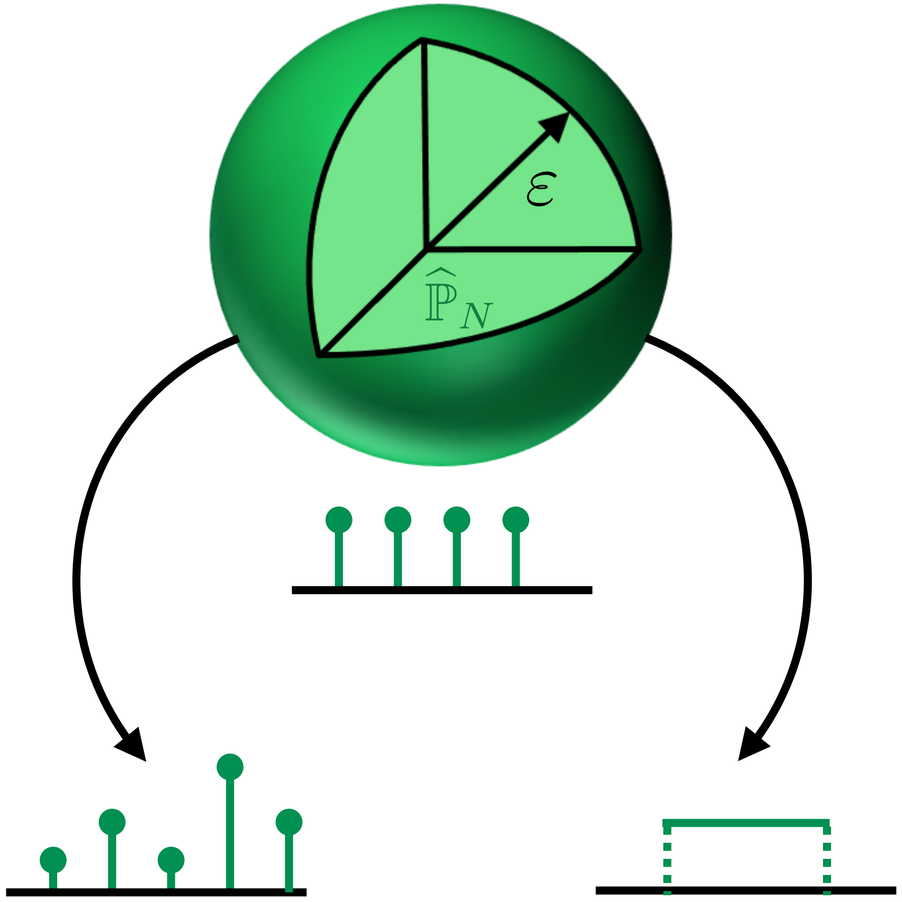
$W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N)$

- discrete OT is easy!
- semi-discrete OT is hard!

[TSK2022]

Optimal Transport Ambiguity Set

$$\mathcal{P} = \left\{ \mathbb{Q} \in \mathcal{M}(\mathbb{R}^n \times \mathbb{Y}) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$$



$$\begin{array}{ll} \sup\limits_{\mathbb{Q} \in \mathcal{M}(\Xi)} & \mathbb{E}_{\mathbb{Q}} \left[L(\theta^\top x, y) \right] \\ \text{s.t.} & W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \end{array}$$

Tractability for Linear Regression

Theorem 1. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If L is convex and Lipschitz, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i) + \varepsilon \mathrm{lip}(L)\|\theta\|_*$$

[SMK15, GCK17, SKM19, BKM19]

Tractability for Linear Regression

Theorem 1. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If L is convex and Lipschitz, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i) + \varepsilon \mathrm{lip}(L)\|\theta\|_*$$

size of ambiguity set

Lipschitz modulus of loss function

dual norm of norm used in c

[SMK15, GCK17, SKM19, BKM19]

Semi-infinite Duality

Lemma 1. Let $\mathcal{P} = \{\mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon\}$. If $c(\xi, \xi) = 0$ for all $\xi \in \Xi$ and $\varepsilon > 0$, then

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}}[I(\xi)] = \inf_{\lambda \geq 0} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi)\right]$$

[MK18, ZG18, BM19, GK16, ZYG22]

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{M}(\Xi)} & \int_{\xi\in\Xi} I(\xi)\mathbb{Q}(\mathrm{d}\xi) \\ \\ \text{s.t.} & W_c(\mathbb{Q},\widehat{\mathbb{P}}_N) \leq \varepsilon \end{cases}$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{M}(\Xi)} & \int_{\xi\in\Xi} I(\xi)\mathbb{Q}(\mathrm{d}\xi) \\ \\ \text{s.t.} & W_c(\mathbb{Q},\widehat{\mathbb{P}}_N) \leq \varepsilon \end{cases}$$

$$= \begin{cases} \sup\limits_{\substack{\mathbb{Q}\in\mathcal{M}(\Xi)\\ \pi\in\mathcal{M}(\Xi\times\Xi)}} & \int_{\xi\in\Xi} I(\xi)\mathbb{Q}(\mathrm{d}\xi) \\ \\ \text{s.t.} & \pi \in \Pi(\mathbb{Q},\widehat{\mathbb{P}}_N) \\ \\ & \int_{\xi\in\Xi}\int_{\xi'\in\Xi} c(\xi,\xi')\pi(\mathrm{d}\xi,\mathrm{d}\xi') \leq \varepsilon \end{cases}$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \left\{ \begin{array}{ll} \sup\limits_{\mathbb{Q}\in\mathcal{M}(\Xi)} & \int_{\xi\in\Xi} I(\xi)\mathbb{Q}(\mathrm{d}\xi) \\ \\ \text{s.t.} & W_c(\mathbb{Q},\widehat{\mathbb{P}}_N) \leq \varepsilon \end{array} \right.$$

$$= \left\{ \begin{array}{ll} \sup\limits_{\substack{\mathbb{Q}\in\mathcal{M}(\Xi)\\ \pi\in\mathcal{M}(\Xi\times\Xi)}} & \int_{\xi\in\Xi} I(\xi)\mathbb{Q}(\mathrm{d}\xi) \\ \\ \text{s.t.} & \pi \in \Pi(\mathbb{Q},\widehat{\mathbb{P}}_N) \\ \\ & \int_{\xi\in\Xi}\int_{\xi'\in\Xi} c(\xi,\xi')\pi(\mathrm{d}\xi,\mathrm{d}\xi') \leq \varepsilon \end{array} \right.$$

$$W_c(\mathbb{Q},\widehat{\mathbb{P}}_N) = \left\{ \begin{array}{ll} \inf\limits_{\pi\in\mathcal{M}(\Xi,\Xi)} & \mathbb{E}_\pi\left[c(\xi,\xi')\right] \\ \text{s.t.} & \pi \in \Pi(\mathbb{Q},\widehat{\mathbb{P}}_N) \end{array} \right.$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \left\{ \begin{array}{ll} \sup\limits_{\mathbb{Q}_i\in\mathcal{M}(\Xi)} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\int_{\xi\in\Xi} I(\xi)\mathbb{Q}_i(\mathrm{d}\xi) \\ & \dfrac{1}{N}\displaystyle\int_{\xi\in\Xi} c(\xi,\widehat{\xi}_i)\mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon \end{array} \right.$$

$$\left\{ \begin{array}{l} \pi\in\Pi(\mathbb{Q},\widehat{\mathbb{P}}_N) \\ \widehat{\mathbb{P}}_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\delta_{\widehat{\xi}_i} \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} \pi = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbb{Q}_i\times\delta_{\widehat{\xi}_i} \\ \mathbb{Q} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbb{Q}_i \end{array} \right.$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \begin{cases} \sup\limits_{\mathbb{Q}_i\in\mathcal{M}(\Xi)} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\int_{\xi\in\Xi} I(\xi)\mathbb{Q}_i(\mathrm{d}\xi) \\ & \dfrac{1}{N}\displaystyle\int_{\xi\in\Xi} c(\xi,\widehat{\xi}_i)\mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon \end{cases}$$

$$\begin{cases} \pi \in \Pi(\mathbb{Q}, \widehat{\mathbb{P}}_N) \\ \widehat{\mathbb{P}}_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\delta_{\widehat{\xi}_i} \end{cases} \Longrightarrow \pi = \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i \times \delta_{\widehat{\xi}_i}$$

$$= \begin{cases} \sup\limits_{\mathbb{Q}_i\geq 0} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\int_{\xi\in\Xi} I(\xi)\mathbb{Q}_i(\mathrm{d}\xi) \\ & \dfrac{1}{N}\displaystyle\int_{\xi\in\Xi} c(\xi,\widehat{\xi}_i)\mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon \\ & \displaystyle\int_{\xi\in\Xi} \mathbb{Q}_i(\mathrm{d}\xi) = 1 \qquad\qquad \forall i \in [N] \end{cases}$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] = \begin{cases} \sup\limits_{\mathbb{Q}_i\in\mathcal{M}(\Xi)} & \dfrac{1}{N}\sum\limits_{i=1}^{N}\int_{\xi\in\Xi} I(\xi)\mathbb{Q}_i(\mathrm{d}\xi) \\ & \dfrac{1}{N}\int_{\xi\in\Xi} c(\xi,\widehat{\xi}_i)\mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon \end{cases}$$

$$\begin{cases} \pi \in \Pi(\mathbb{Q}, \widehat{\mathbb{P}}_N) \\ \widehat{\mathbb{P}}_N = \dfrac{1}{N}\sum\limits_{i=1}^{N}\delta_{\widehat{\xi}_i} \end{cases} \Longrightarrow \pi = \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i \times \delta_{\widehat{\xi}_i}$$

$$= \begin{cases} \sup\limits_{\mathbb{Q}_i\geq 0} & \dfrac{1}{N}\sum\limits_{i=1}^{N}\int_{\xi\in\Xi} I(\xi)\mathbb{Q}_i(\mathrm{d}\xi) \\ & \dfrac{1}{N}\int_{\xi\in\Xi} c(\xi,\widehat{\xi}_i)\mathbb{Q}_i(\mathrm{d}\xi) \leq \varepsilon \qquad (\lambda) \\ & \int_{\xi\in\Xi}\mathbb{Q}_i(\mathrm{d}\xi) = 1 \qquad (s_i) \qquad \forall i \in [N] \end{cases}$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] \leq \begin{cases} \inf\limits_{\substack{\lambda\geq 0 \\ s\in\mathbb{R}^N}} & \lambda\varepsilon + \sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \dfrac{\lambda}{N} c(\xi, \widehat{\xi}_i) + s_i \geq \dfrac{1}{N} I(\xi) \quad \forall \xi \in \Xi \end{cases}$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] \leq \begin{cases} \inf\limits_{\substack{\lambda\geq 0 \\ s\in\mathbb{R}^N}} & \lambda\varepsilon + \sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \frac{\lambda}{N} c(\xi, \widehat{\xi}_i) + s_i \geq \frac{1}{N} I(\xi) \quad \forall \xi \in \Xi \end{cases}$$

$$= \inf_{\lambda\geq 0} \ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi\in\Xi} \ I(\xi) - \lambda c(\xi, \widehat{\xi}_i)$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] \leq \begin{cases} \inf\limits_{\substack{\lambda\geq 0 \\ s\in\mathbb{R}^N}} & \lambda\varepsilon + \sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \frac{\lambda}{N} c(\xi,\widehat{\xi}_i) + s_i \geq \frac{1}{N} I(\xi) \quad \forall \xi \in \Xi \end{cases}$$

$$= \inf_{\lambda\geq 0} \ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi\in\Xi} \ I(\xi) - \lambda c(\xi,\widehat{\xi}_i)$$

$$= \inf_{\lambda\geq 0} \ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta\in\Xi} I(\zeta) - \lambda c(\zeta,\xi)\right]$$

Proof of Lemma 1

$$\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[I(\xi)\right] \overset{\text{[S01]}}{\leq} \begin{cases} \inf\limits_{\substack{\lambda\geq 0 \\ s\in\mathbb{R}^N}} & \lambda\varepsilon + \sum\limits_{i=1}^{N} s_i \\ \text{s.t.} & \frac{\lambda}{N} c(\xi, \widehat{\xi}_i) + s_i \geq \frac{1}{N} I(\xi) \quad \forall \xi \in \Xi \end{cases}$$

$$= \inf_{\lambda\geq 0} \ \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi\in\Xi} \ I(\xi) - \lambda c(\xi, \widehat{\xi}_i)$$

$$= \inf_{\lambda\geq 0} \ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta\in\Xi} I(\zeta) - \lambda c(\zeta, \xi)\right]$$

Semi-infinite Duality

Lemma 1. Let $\mathcal{P} = \{\mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon\}$. If $c(\xi, \xi) = 0$ for all $\xi \in \Xi$ and $\varepsilon > 0$, then

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}}[I(\xi)] = \inf_{\lambda \geq 0} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi)\right]$$

[ZG18, PK18, BM19, GK16]

Lipschitz Envelope

Lemma 2. Let $L : \mathbb{R} \to \mathbb{R}$ be a convex and Lipschitz function. Then,

$$\sup_{\zeta \in \mathbb{R}^n} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \begin{cases} L(\theta^\top \xi + \theta_0) & \text{if } \; \mathrm{lip(L)} \|\theta\|_* \leq \lambda \\ +\infty & \text{else} \end{cases}$$

for any $\theta, \xi \in \mathbb{R}^n, \theta_0 \in \mathbb{R}$ and $\lambda > 0$.

[SKM19]

Proof of Lemma 2

$$L(\theta^\top \zeta + \theta_0) = L^{**}(\theta^\top \zeta + \theta_0) = \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa)$$

$$\mathcal{K} = \{\kappa \in \mathbb{R} : L^*(\kappa) < \infty\}$$

Proof of Lemma 2

$$L(\theta^\top \zeta + \theta_0) = L^{**}(\theta^\top \zeta + \theta_0) = \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa)$$

$$\mathcal{K} = \{\kappa \in \mathbb{R} : L^*(\kappa) < \infty\}$$

$$\sup_{\zeta} \ L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\zeta} \ \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - \lambda \|\zeta - \xi\|$$

Proof of Lemma 2

$$L(\theta^\top \zeta + \theta_0) = L^{**}(\theta^\top \zeta + \theta_0) = \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa)$$

$$\mathcal{K} = \{\kappa \in \mathbb{R} : L^*(\kappa) < \infty\}$$

$$\sup_{\zeta} \ L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\zeta} \ \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - \boxed{\lambda \|\zeta - \xi\|}$$

$$= \sup_{\kappa \in \mathcal{K}} \ \sup_{\zeta} \ \inf_{\|p\|_* \leq \lambda} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - p^\top (\zeta - \xi)$$

$$\sup_{\|p\|_* \leq \lambda} \ p^\top (\zeta - \xi)$$

Proof of Lemma 2

$$L(\theta^\top \zeta + \theta_0) = L^{**}(\theta^\top \zeta + \theta_0) = \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa)$$

$$\mathcal{K} = \{\kappa \in \mathbb{R} : L^*(\kappa) < \infty\}$$

$$\begin{aligned}
\sup_{\zeta} \ L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| &= \sup_{\zeta} \ \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - \lambda \|\zeta - \xi\| \\
&= \sup_{\kappa \in \mathcal{K}} \ \sup_{\zeta} \ \inf_{\|p\|_* \leq \lambda} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - p^\top (\zeta - \xi) \\
\textbf{(Sion's minimax)} \quad &= \sup_{\kappa \in \mathcal{K}} \ \inf_{\|p\|_* \leq \lambda} \ \sup_{\zeta} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - p^\top (\zeta - \xi)
\end{aligned}$$

Proof of Lemma 2

$$L(\theta^\top \zeta + \theta_0) = L^{**}(\theta^\top \zeta + \theta_0) = \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa)$$

$$\mathcal{K} = \{\kappa \in \mathbb{R} : L^*(\kappa) < \infty\}$$

$$\begin{aligned}
\sup_{\zeta} \ L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| &= \sup_{\zeta} \ \sup_{\kappa \in \mathcal{K}} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - \lambda \|\zeta - \xi\| \\
&= \sup_{\kappa \in \mathcal{K}} \ \sup_{\zeta} \ \inf_{\|p\|_* \le \lambda} \ \kappa(\theta^\top \zeta + \theta_0) - L^*(\kappa) - p^\top (\zeta - \xi) \\
\textbf{(Sion's minimax)} \quad &= \sup_{\kappa \in \mathcal{K}} \ \inf_{\|p\|_* \le \lambda} \ \kappa\theta_0 - L^*(\kappa) + p^\top \xi + \begin{cases} 0 & \text{if } \kappa\theta - p = 0 \\ +\infty & \text{else} \end{cases}
\end{aligned}$$

Proof of Lemma 2

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\kappa \in \mathcal{K}} \begin{cases} \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

Proof of Lemma 2

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\kappa \in \mathcal{K}} \begin{cases} \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} \sup_{\kappa \in \mathcal{K}} \; \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \; \forall \kappa \in \mathcal{K} \\ \\ +\infty & \text{else} \end{cases}$$

Proof of Lemma 2

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\kappa \in \mathcal{K}} \begin{cases} \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} \sup\limits_{\kappa \in \mathcal{K}} \; \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \; \forall \kappa \in \mathcal{K} \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} \sup\limits_{\kappa \in \mathcal{K}} \; \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \sup\limits_{\kappa \in \mathcal{K}} \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

Proof of Lemma 2

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \sup_{\kappa \in \mathcal{K}} \begin{cases} \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} \sup_{\kappa \in \mathcal{K}} \; \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \|\kappa\theta\|_* \leq \lambda \; \forall \kappa \in \mathcal{K} \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} \sup_{\kappa \in \mathcal{K}} \; \kappa(\theta^\top \xi + \theta_0) - L^*(\kappa) & \text{if } \sup_{\kappa \in \mathcal{K}} \|\kappa\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

$$= \begin{cases} L(\theta^\top \xi + \theta_0) & \text{if } \mathrm{lip}(L)\|\theta\|_* \leq \lambda \\ \\ +\infty & \text{else} \end{cases}$$

Tractability for Linear Regression

Theorem 1. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If L is convex and Lipschitz, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i) + \varepsilon \mathrm{lip}(L)\|\theta\|_*$$

[SMK15, GCK17, SKM19, BKM19]

Proof of Theorem 1

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x' - y') - \lambda\|x' - x\| - \lambda\delta_{y'=y}\right]$$

Semi-infinite Duality

Lemma 1. Let $\mathcal{P} = \{\mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon\}$. If $c(\xi, \xi) = 0$ for all $\xi \in \Xi$ and $\varepsilon > 0$, then

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}}[I(\xi)] = \inf_{\lambda \geq 0} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi)\right]$$

[ZG18, PK18, BM19, GK16]

Proof of Theorem 1

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x' - y') - \lambda\|x' - x\| - \lambda\delta_{y'=y}\right]$$

$$= \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x' - y) - \lambda\|x' - x\|\right]$$

Proof of Theorem 1

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x-y)\right]$$

$$\textbf{(Lemma 1)}=\inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x'-y')-\lambda\|x'-x\|-\lambda\delta_{y'=y}\right]$$

$$=\inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x'-y)-\lambda\|x'-x\|\right]$$

$$\textbf{(Lemma 2)}=\inf_{\substack{\theta\in\Theta\\ \lambda\geq \mathrm{lip}(L)\|\theta\|_*}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]$$

Lipschitz Envelope

Lemma 2. Let $L : \mathbb{R} \to \mathbb{R}$ be a convex and Lipschitz function. Then,

$$\sup_{\zeta \in \mathbb{R}^n} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\| = \begin{cases} L(\theta^\top \xi + \theta_0) & \text{if } \; \mathrm{lip(L)} \|\theta\|_* \leq \lambda \\ +\infty & \text{else} \end{cases}$$

for any $\theta, \xi \in \mathbb{R}^n, \theta_0 \in \mathbb{R}$ and $\lambda > 0$.

[SKM19]

Proof of Theorem 1

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x-y)\right]$$

$$\textbf{(Lemma 1)}=\inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x'-y')-\lambda\|x'-x\|-\lambda\delta_{y'=y}\right]$$

$$=\inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x'-y)-\lambda\|x'-x\|\right]$$

$$\textbf{(Lemma 2)}=\inf_{\substack{\theta\in\Theta\\ \lambda\geq \mathrm{lip}(L)\|\theta\|_*}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]$$

$$=\inf_{\theta\in\Theta}\ \varepsilon\mathrm{lip}(L)\|\theta\|_*+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]$$

Tractability for Linear Regression

Theorem 1. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If L is convex and Lipschitz, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i) + \varepsilon \mathrm{lip}(L)\|\theta\|_*$$

[SMK15, GCK17, SKM19, BKM19]

Examples

Robust Regression

$$L(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \delta \\ \delta\left(|z| - \frac{1}{2}\delta\right) & \text{else} \end{cases}$$

$$\text{lip}(L) = \delta$$

Support Vector Regression

$$L(z) = \max\{0, |z| - \varepsilon\}$$

$$\text{lip}(L) = 1$$

Quantile Regression

$$L(z) = \max\{-\tau z, (1-\tau)z\}$$

$$\text{lip}(L) = \max\{\tau, 1-\tau\}$$

Examples

Least Square Regression

$$L(z) = \frac{1}{2}z^2$$

$$\mathrm{lip}(L) = \infty$$

Least Squares Regression

Theorem 2. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\|^2 + \delta_{y=y'}$.

If $L = z^2$, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta} \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i)} + \sqrt{\varepsilon}\|\theta\|_*\right)^2$$

[BKM19]

Semi-infinite Duality

Lemma 1. Let $\mathcal{P} = \{\mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon\}$. If $c(\xi, \xi) = 0$ for all $\xi \in \Xi$ and $\varepsilon > 0$, then

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}}[I(\xi)] = \inf_{\lambda \geq 0} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi)\right]$$

[ZG18, PK18, BM19, GK16]

Moreau Envelope

Lemma 3. Let $L(z) = z^2$. Then,

$$\sup_{\zeta \in \mathbb{R}^n} L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 = \begin{cases} \dfrac{\lambda}{\lambda - \|\theta\|_*^2} L(\theta^\top \xi + \theta_0) & \text{if } \|\theta\|_*^2 < \lambda \\ +\infty & \text{else} \end{cases}$$

for any $\theta, \xi \in \mathbb{R}^n, \theta_0 \in \mathbb{R}$ and $\lambda > 0$.

[BKM19, SADK22]

Proof of Lemma 3

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2$$

$$= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2$$

Proof of Lemma 3

$$
\begin{aligned}
&\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 \\
&= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right.
\end{aligned}
$$

Proof of Lemma 3

$$
\begin{aligned}
&\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 \\
&= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} \quad &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 - \kappa\gamma + \kappa\theta^\top \Delta
\end{aligned}
$$

Proof of Lemma 3

$$
\begin{aligned}
&\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 \\
&= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 - \kappa\gamma + \kappa\theta^\top \Delta \\
\textbf{(Holder inequality)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 - \kappa\gamma + \|\kappa\theta^\top\|_* \|\Delta\|
\end{aligned}
$$

Proof of Lemma 3

$$
\begin{aligned}
&\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 \\
&= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 - \kappa\gamma + \kappa\theta^\top \Delta \\
\textbf{(Holder inequality)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 - \kappa\gamma + \|\kappa\theta^\top\|_* \|\Delta\| \\
&= \sup_{\gamma} \; \inf_{\kappa} \; L(\gamma + \theta^\top \xi + \theta_0) - \kappa\gamma + \frac{\|\kappa\theta\|_*^2}{4\lambda}
\end{aligned}
$$

Proof of Lemma 3

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 = \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda \gamma^2}{\|\theta\|_*^2}$$

Proof of Lemma 3

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 = \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda \gamma^2}{\|\theta\|_*^2}$$

$$= \sup_{\gamma} \; (\gamma + \theta^\top \xi)^2 - \frac{\lambda \gamma^2}{\|\theta\|_*^2}$$

Proof of Lemma 3

$$\sup_{\zeta} \; L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 = \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda \gamma^2}{\|\theta\|_*^2}$$

$$= \sup_{\gamma} \; (\gamma + \theta^\top \xi)^2 - \frac{\lambda \gamma^2}{\|\theta\|_*^2}$$

$$= \begin{cases} \dfrac{\lambda}{\lambda - \|\theta\|_*^2} (\theta^\top \xi)^2 & \text{if} \quad \|\theta\|_*^2 < \lambda \\ +\infty & \text{else} \end{cases}$$

Least Squares Regression

Theorem 2. Suppose that $c\big((x,y),(x',y')\big) = \|x-x'\|^2 + \delta_{y=y'}$.

If $L = z^2$, then

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right] = \inf_{\theta\in\Theta}\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} L(\theta^\top \widehat{x}_i - \widehat{y}_i)} + \sqrt{\varepsilon}\|\theta\|_*\right)^2$$

[BKM19]

Proof of Theorem 2

$$\inf_{\theta\in\Theta}\ \sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x' - y') - \lambda\|x'-x\|^2 - \lambda\delta_{y'=y}\right]$$

Proof of Theorem 2

$$\inf_{\theta\in\Theta}\ \sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x - y)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x' - y') - \lambda\|x'-x\|^2 - \lambda\delta_{y'=y}\right]$$

$$= \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x' - y) - \lambda\|x'-x\|^2\right]$$

Proof of Theorem 2

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x-y)\right]$$

$$\textbf{(Lemma 1)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x'-y')-\lambda\|x'-x\|^2-\lambda\delta_{y'=y}\right]$$

$$=\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x'-y)-\lambda\|x'-x\|^2\right]$$

$$\textbf{(Lemma 3)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda>\|\theta\|_*^2}}\ \lambda\varepsilon+\frac{\lambda}{\lambda-\|\theta\|_*^2}\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]$$

Proof of Theorem 2

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(\theta^\top x-y)\right]$$

$$\textbf{(Lemma 1)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(\theta^\top x'-y')-\lambda\|x'-x\|^2-\lambda\delta_{y'=y}\right]$$

$$=\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(\theta^\top x'-y)-\lambda\|x'-x\|^2\right]$$

$$\textbf{(Lemma 3)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda>\|\theta\|_*^2}}\ \lambda\varepsilon+\frac{\lambda}{\lambda-\|\theta\|_*^2}\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]$$

$$=\ \inf_{\theta\in\Theta}\ \left(\sqrt{\varepsilon}\|\theta\|_*+\sqrt{\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(\theta^\top x-y)\right]}\right)^2$$

Tractability for Linear Classification

Theorem 3. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If L is convex and Lipschitz, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[yL(\theta^\top x)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L(\widehat{y}_i\theta^\top\widehat{x}_i) + \varepsilon \mathrm{lip}(L)\|\theta\|_*$$

[SMK15, GCK17, SKM19, BKM19]

Proof of Theorem 3

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]$$

$$\textbf{(Lemma 1)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x')-\lambda\|x'-x\|-\lambda\delta_{y'=y}\right]$$

Proof of Theorem 3

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]$$

$$\textbf{(Lemma 1)}\ =\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x')-\lambda\|x'-x\|-\lambda\delta_{y'=y}\right]$$

$$=\ \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon+\mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x')-\lambda\|x'-x\|\right]$$

Proof of Theorem 3

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x') - \lambda\|x'-x\| - \lambda\delta_{y'=y}\right]$$

$$= \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x') - \lambda\|x'-x\|\right]$$

$$\textbf{(Lemma 2)} = \inf_{\substack{\theta\in\Theta\\ \lambda\geq \mathrm{lip}(L)\|\theta\|_*}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(y\theta^\top x)\right]$$

Proof of Theorem 3

$$
\begin{aligned}
&\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{P}}\ \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]\\
\textbf{(Lemma 1)}\ &= \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x') - \lambda\|x'-x\| - \lambda\delta_{y'=y}\right]\\
&= \inf_{\substack{\theta\in\Theta\\ \lambda\geq 0}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x') - \lambda\|x'-x\|\right]\\
\textbf{(Lemma 2)}\ &= \inf_{\substack{\theta\in\Theta\\ \lambda\geq \mathrm{lip}(L)\|\theta\|_*}}\ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(y\theta^\top x)\right]\\
&= \inf_{\theta\in\Theta}\ \varepsilon\mathrm{lip}(L)\|\theta\|_* + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L(y\theta^\top x)\right]
\end{aligned}
$$

Examples

Support Vector Machine

$$L(z) = \max\{0, 1 - z\}$$

$$\operatorname{lip}(L) = 1$$

Support Vector Machine II

$$L(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1-z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{else} \end{cases}$$

$$\operatorname{lip}(L) = 1$$

Logistic Regression

$$L(z) = \log(1 + \exp(-z))$$

$$\operatorname{lip}(L) = 1$$

Examples

Ideal Classification

$$L(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 0 & \text{else} \end{cases}$$

nonconvex!

Ideal Classification

Theorem 4. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If $L(z) = \mathbb{1}_{z\leq 0}$, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L_R(\widehat{y}_i\theta^\top \widehat{x}_i) + \varepsilon\|\theta\|_*$$

where $L_R(z) = \max\{0, 1-z\} + \max\{0, -z\}$.

[H-NW22]

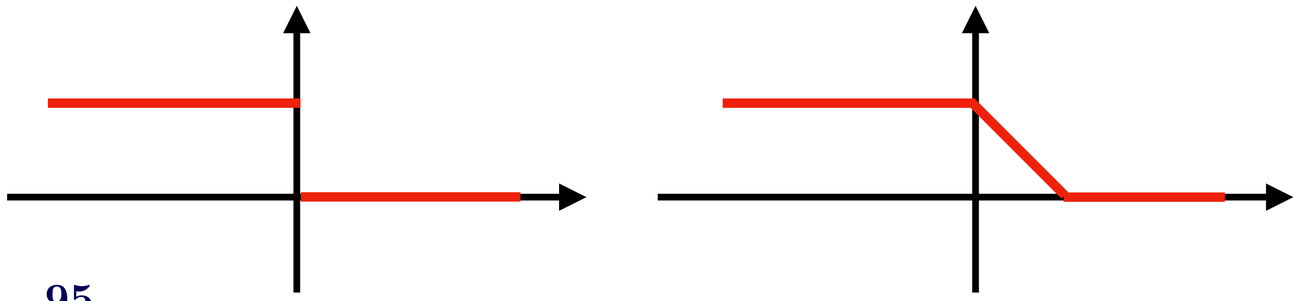
Ideal Classification

Theorem 4. Suppose that $c\big((x,y),(x',y')\big) = \|x - x'\| + \delta_{y=y'}$.

If $L(z) = \mathbb{1}_{z\leq 0}$, then

$$\inf_{\theta\in\Theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right] = \inf_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} L_R(\widehat{y}_i\theta^\top \widehat{x}_i) + \varepsilon\|\theta\|_*$$

where $L_R(z) = \max\{0, 1-z\} + \max\{0, -z\}$.



[H-NW22]

Lipschitz Envelope II

Lemma 4. Let $L(z) = \mathbb{1}_{z \leq 0}$. Then,

$$\sup_{\zeta \in \mathbb{R}^n} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = L_R\big(\lambda(\theta^\top \xi)/\|\theta\|_*\big)$$

for any $\theta, \xi \in \mathbb{R}^n$ and $\lambda > 0$.

[SADK22]

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\|$$

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\|$$

$$= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right.$$

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\|$$

$$= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right.$$

$$\textbf{(Slater condition)} = \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| - \kappa\gamma + \kappa\theta^\top \Delta$$

Proof of Lemma 4

$$
\begin{aligned}
\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| &= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\| \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| - \kappa\gamma + \kappa\theta^\top \Delta \\
&= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; \inf_{\|p\| \le \lambda} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa\gamma + \kappa\theta^\top \Delta
\end{aligned}
$$

Proof of Lemma 4

$$
\begin{aligned}
\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| &= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\| \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| - \kappa\gamma + \kappa\theta^\top \Delta \\
&= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; \inf_{\|p\| \le \lambda} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa\gamma + \kappa\theta^\top \Delta \\
\textbf{(Sion's minimax)} &= \sup_{\gamma} \; \inf_{\kappa} \; \inf_{\|p\| \le \lambda} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa\gamma + \kappa\theta^\top \Delta
\end{aligned}
$$

Proof of Lemma 4

$$
\begin{aligned}
\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| &= \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\| \\
&= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right. \\
\textbf{(Slater condition)} &= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| - \kappa \gamma + \kappa \theta^\top \Delta \\
&= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; \inf_{\|p\| \le \lambda} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa \gamma + \kappa \theta^\top \Delta \\
\textbf{(Sion's minimax)} &= \sup_{\gamma} \; \inf_{\|p\| \le \lambda} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa \gamma + \kappa \theta^\top \Delta \\
&= \sup_{\gamma} \; \inf_{\|\kappa \theta\|_* \le \lambda} \; L(\gamma + \theta^\top \xi) - \kappa \gamma
\end{aligned}
$$

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\Delta} \; L(\theta^\top \Delta + \theta^\top \xi) - \lambda \|\Delta\|$$

$$= \sup_{\gamma} \left\{ \begin{array}{ll} \sup\limits_{\Delta} & L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| \\ \text{s.t.} & \gamma = \theta^\top \Delta \end{array} \right.$$

$$\textbf{(Slater condition)} = \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; L(\gamma + \theta^\top \xi) - \lambda \|\Delta\| - \kappa\gamma + \kappa\theta^\top \Delta$$

$$= \sup_{\gamma} \; \inf_{\kappa} \; \sup_{\Delta} \; \inf_{\|p\| \le \lambda} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa\gamma + \kappa\theta^\top \Delta$$

$$\textbf{(Sion's minimax)} = \sup_{\gamma} \; \inf_{\|p\| \le \lambda} \; \sup_{\kappa, \Delta} \; L(\gamma + \theta^\top \xi) - p^\top \Delta - \kappa\gamma + \kappa\theta^\top \Delta$$

$$= \sup_{\gamma} \; \inf_{\|\kappa\theta\|_* \le \lambda} \; L(\gamma + \theta^\top \xi) - \kappa\gamma$$

$$\kappa^\star = \frac{\lambda}{\|\theta\|_*} \operatorname{sgn}(\gamma)$$

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda |\gamma|}{\|\theta\|_*}$$

Proof of Lemma 4

$$\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| = \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda|\gamma|}{\|\theta\|_*}$$

$$= \sup_{\gamma} \; L(\gamma\|\theta\|_* + \theta^\top \xi) - \lambda|\gamma|$$

Proof of Lemma 4

$$\begin{aligned}
\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| &= \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda|\gamma|}{\|\theta\|_*} \\
&= \sup_{\gamma} \; L(\gamma\|\theta\|_* + \theta^\top \xi) - \lambda|\gamma| \\
&= \begin{cases} 1 & \text{if } \theta^\top \xi \leq 0 \\ \max\{0, 1 - \lambda \frac{\theta^\top \xi}{\|\theta\|_*}\} & \text{else} \end{cases}
\end{aligned}$$

Proof of Lemma 4

$$
\begin{aligned}
\sup_{\zeta} \; L(\theta^\top \zeta) - \lambda \|\zeta - \xi\| &= \sup_{\gamma} \; L(\gamma + \theta^\top \xi) - \frac{\lambda|\gamma|}{\|\theta\|_*} \\
&= \sup_{\gamma} \; L(\gamma\|\theta\|_* + \theta^\top \xi) - \lambda|\gamma| \\
&= \begin{cases} 1 & \text{if } \theta^\top \xi \leq 0 \\ \max\{0, 1 - \lambda \frac{\theta^\top \xi}{\|\theta\|_*}\} & \text{else} \end{cases} \\
&= L_R\big(\lambda(\theta^\top \xi)/\|\theta\|_*\big)
\end{aligned}
$$

Proof of Theorem 4

$$\inf_{\theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[L(y \theta^\top x) \right]$$

$$\textbf{(Lemma 1)} \quad = \inf_{\substack{\theta \\ \lambda \geq 0}} \lambda \varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N} \left[\sup_{(x', y') \in \mathbb{R}^n \times \mathbb{R}} L(y' \theta^\top x') - \lambda \| x' - x \| - \lambda \delta_{y' = y} \right]$$

Proof of Theorem 4

$$\inf_{\theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[L(y \theta^\top x) \right]$$

$$\textbf{(Lemma 1)} \quad = \inf_{\substack{\theta \\ \lambda \geq 0}} \lambda \varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N} \left[\sup_{(x', y') \in \mathbb{R}^n \times \mathbb{R}} L(y' \theta^\top x') - \lambda \| x' - x \| - \lambda \delta_{y' = y} \right]$$

$$= \inf_{\substack{\theta \\ \lambda \geq 0}} \lambda \varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N} \left[\sup_{x' \in \mathbb{R}^n} L(y \theta^\top x') - \lambda \| x' - x \| \right]$$

Proof of Theorem 4

$$
\begin{aligned}
& \inf_{\theta} \sup_{\mathbb{Q}\in\mathcal{P}} \ \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right] \\
\textbf{(Lemma 1)} \quad &= \inf_{\substack{\theta \\ \lambda\geq 0}} \ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x') - \lambda\|x'-x\| - \lambda\delta_{y'=y}\right] \\
&= \inf_{\substack{\theta \\ \lambda\geq 0}} \ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x') - \lambda\|x'-x\|\right] \\
\textbf{(Lemma 4)} \quad &= \inf_{\substack{\theta \\ \lambda\geq 0}} \ \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L_R\big(\lambda(y\theta^\top x)/\|\theta\|_*\big)\right]
\end{aligned}
$$

Proof of Theorem 4

$$\inf_{\theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]$$

$$\textbf{(Lemma 1)} \quad = \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x') - \lambda\|x'-x\| - \lambda\delta_{y'=y}\right]$$

$$= \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x') - \lambda\|x'-x\|\right]$$

$$\textbf{(Lemma 4)} \quad = \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L_R\big(\lambda(y\theta^\top x)/\|\theta\|_*\big)\right]$$

$$\theta \leftarrow \lambda\theta/\|\theta\|_* \Downarrow \lambda = \|\theta\|_*$$

Proof of Theorem 4

$$\inf_{\theta} \sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\mathbb{Q}}\left[L(y\theta^\top x)\right]$$

$$\textbf{(Lemma 1)} = \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{(x',y')\in\mathbb{R}^n\times\mathbb{R}} L(y'\theta^\top x') - \lambda\|x'-x\| - \lambda\delta_{y'=y}\right]$$

$$= \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[\sup_{x'\in\mathbb{R}^n} L(y\theta^\top x') - \lambda\|x'-x\|\right]$$

$$\textbf{(Lemma 4)} = \inf_{\substack{\theta \\ \lambda\geq 0}} \lambda\varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L_R\big(\lambda(y\theta^\top x)/\|\theta\|_*\big)\right]$$

$$= \inf_{\theta} \varepsilon\|\theta\|_* + \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[L_R(y\theta^\top x)\right]$$

$$\theta \leftarrow \lambda\theta/\|\theta\|_* \Downarrow \lambda = \|\theta\|_*$$

Conclusion

$$\textbf{Regularization} = \textbf{Distributional Robustness}$$

Take Away

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right]$$

$$\mathcal{P} = \{ \mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \mathbb{P}) \leq \varepsilon \}$$

Take Away

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right]$$

$$\mathcal{P} = \{ \mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \mathbb{P}) \leq \varepsilon \}$$

Step 1

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right] = \inf_{\lambda \geq 0} \lambda \varepsilon + \mathbb{E}_{\mathbb{P}} \left[\sup_{\zeta \in \Xi} \ell(\theta, \zeta) - \lambda c(\zeta, \xi) \right]$$

[MK18, ZG18, BM19, GK16, ZYG22]

Take Away

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right]$$

$$\mathcal{P} = \{ \mathbb{Q} \in \mathcal{M}(\Xi) : W_c(\mathbb{Q}, \mathbb{P}) \leq \varepsilon \}$$

Step 1

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} \left[\ell(\theta, \xi) \right] = \inf_{\lambda \geq 0} \lambda \varepsilon + \mathbb{E}_{\mathbb{P}} \left[\sup_{\zeta \in \Xi} \ell(\theta, \zeta) - \lambda c(\zeta, \xi) \right]$$

Step 2

References

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative ad- versarial networks. In International Conference on Machine Learning, pages 214–223, 2017.

[BK16] J. Blanchet and Y. Kang. Sample out-of-sample inference based on Wasserstein distance. arXiv:1605.01340, 2016.

[BKM19] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. Journal of Applied Probability, 56(3):830–857, 2019.

[BM19] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. Mathematics of Operations Research, 44(2):565– 600, 2019.

[BMN21] J. Blanchet, K. Murthy, and V. A. Nguyen. Statistical Analysis of Wasserstein Distributionally Robust Estimators. arXiv:2108.02120, 2021.

[BMZ18] J. Blanchet, K. Murthy, and F. Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. arXiv:1810.02403, 2018.

[CP18] R. Chen and I. C. Paschalidis. A robust learning approach for regres- sion models based on distributionally robust optimization. Journal of Machine Learning Research, 19(1):517–564, 2018.

[Gal16] A. Galichon. Optimal transport methods in economics. Princeton University Press, 2016.

[Gao20] R. Gao. Finite-Sample Guarantees for Wasserstein Distribution- ally Robust Optimization: Breaking the Curse of Dimensionality. arXiv:2009.04382, 2020. [GCK20] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. arXiv:1712.06050v3, 2020.

[GK17] R. Gao and A. J. Kleywegt. Distributionally robust stochastic opti- mization with dependence structure. arXiv:1701.04200, 2017.

[HGK+16] G. Huang, C. Guo, M. J. Kusner, Y. Sun, F. Sha, and K. Q. Wein- berger. Supervised Word Mover's Distance. In Advances in Neural Information Processing Systems, pages 4862–4870, 2016.

[HNW20] N. Ho-Nguyen and S. J. Wright. Adversarial Classification via Distri- butional Robustness with Wasserstein Ambiguity. arXiv:2005.13815, 2020. [

LS18] B. Levy and E. L. Schwindt. Notions of optimal transport theory and how to implement them on a computer. Computers & Graphics, 72:135–148, 2018.

[MK18] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming, 171(1- 2):115–166, 2018.

[PPC10] N. Papadakis, E. Provenzi, and V. Caselles. A variational model for histogram transfer of color images. IEEE Transactions on Image Processing, 20(6):1682–1695, 2010.

[RBHdM19] H. Rahimian, G. Bayraksan, and T. Homem-de Mello. Identifying effective scenarios in distributionally robust stochastic programs with total variation distance.

[SKM19] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via Mass Transportation. Journal of Machine Learning Research, 20(103):1–68, 2019.

[SMK15] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In Advances in Neural Information Processing Systems, pages 1576–1584, 2015. [

TPJR18] Y. Tian, K. Pei, S. Jana, and B. Ray. Deeptest: Automated test- ing of deep-neural-network-driven autonomous cars. In International Conference on Software Engineering, pages 303–314, 2018.

[ZG18] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. Operations Research Letters, 46(2):262–267, 2018.

[ZYG22] L. Zhang, J. Yang, and R. Gao. A Simple Duality Proof for Wasserstein Distributionally Robust Optimization. arXiv preprint arXiv:2205.00362, 2022.