

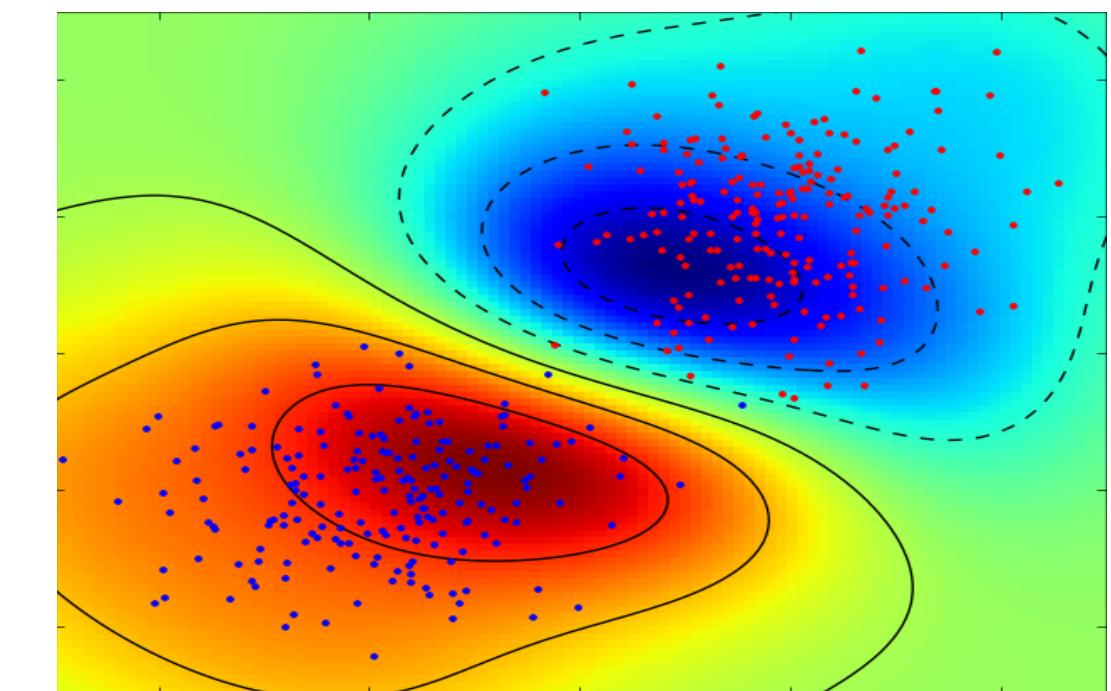
# Regularization via Optimal Transport

Soroosh Shafieezadeh Abadeh, Tepper School of Business, CMU

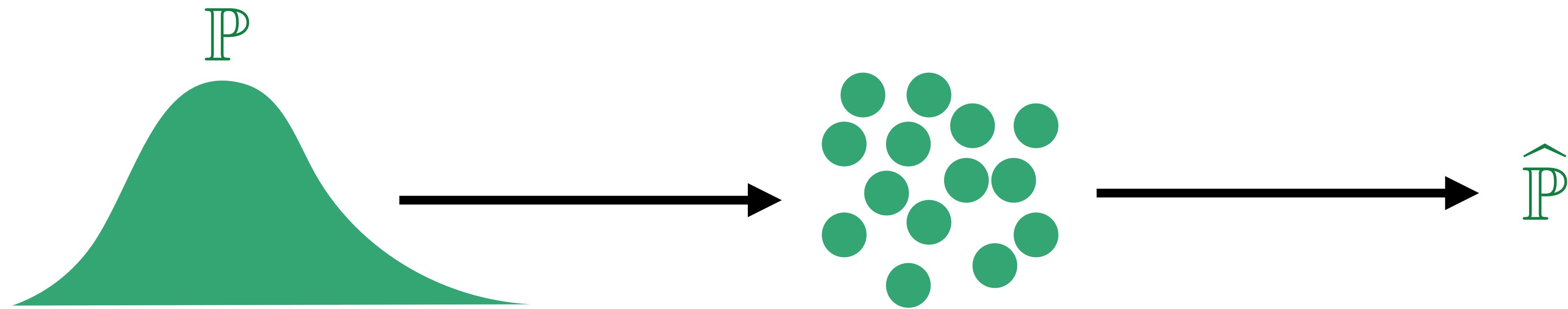


# Stochastic Programming

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



# Failure Examples: Overfitting



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

# Failure Examples: Adversarial Attack [GSS15]



$+ .007 \times$



=



Panda  
57.7% Confidence

Gibbon  
99.3% Confidence

# Failure Examples: Fake Data

THE WALL STREET JOURNAL.

[Subscribe](#) | [Sign In](#)

[Home](#) [World](#) [U.S.](#) [Politics](#) [Economy](#) [Business](#) [Tech](#) [Markets](#) [Opinion](#) [Books & Arts](#) [Real Estate](#) [Life & Work](#) [WSJ. Magazine](#) [Sports](#) 

TECH | PERSONAL TECH | PERSONAL TECHNOLOGY: NICOLE NGUYEN

SHARE



## 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



PRINT



TEXT

191



 Listen to article (10 minutes)

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

### UPCOMING EVENTS



Oct

5

2021

12:00 PM - 5:00 PM EDT

WSJ Jobs Summit

Oct

6

2021

12:30 PM - 2:00 PM EDT

The Future Of Health

# Failure Examples: Domain Change [TPJR18]



original



fog



original



rain



original



shear(0.1)



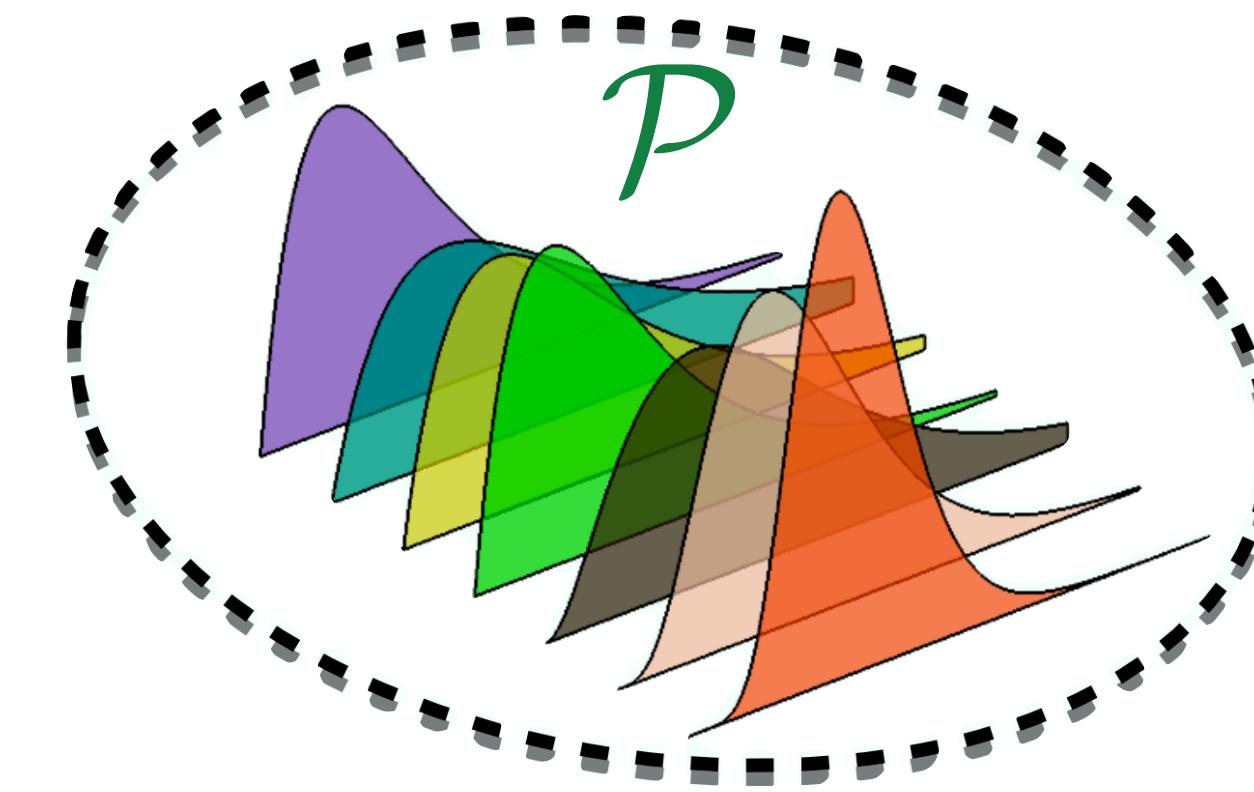
original



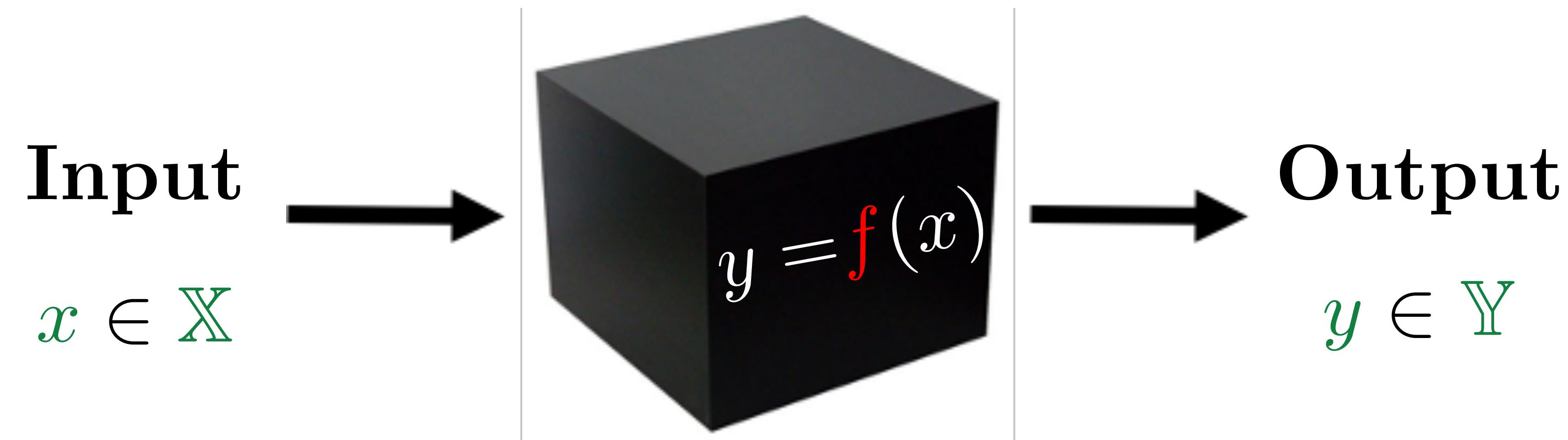
rotation(6 degree)

# Distributionally Robust Optimization

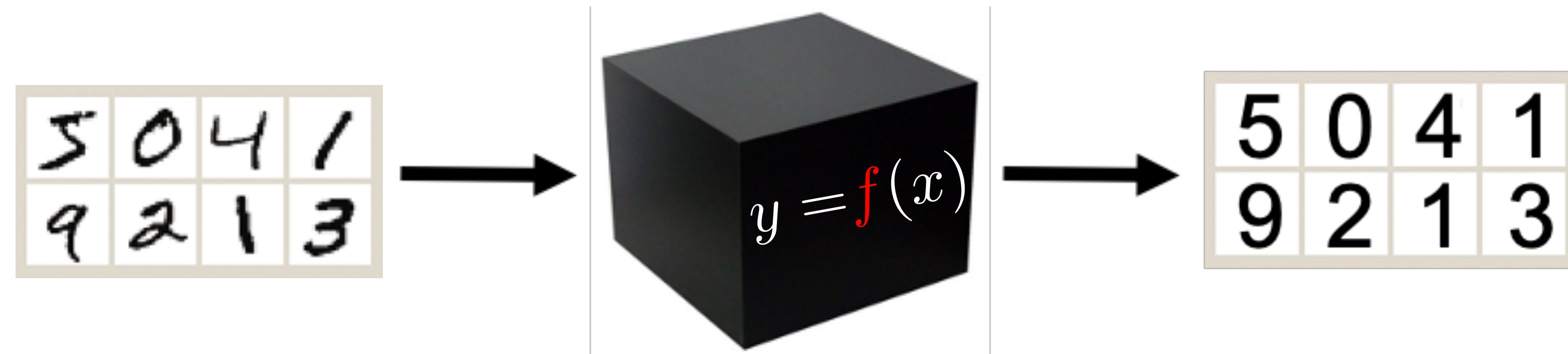
$$\inf_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_Q [\ell(\theta, \xi)]$$



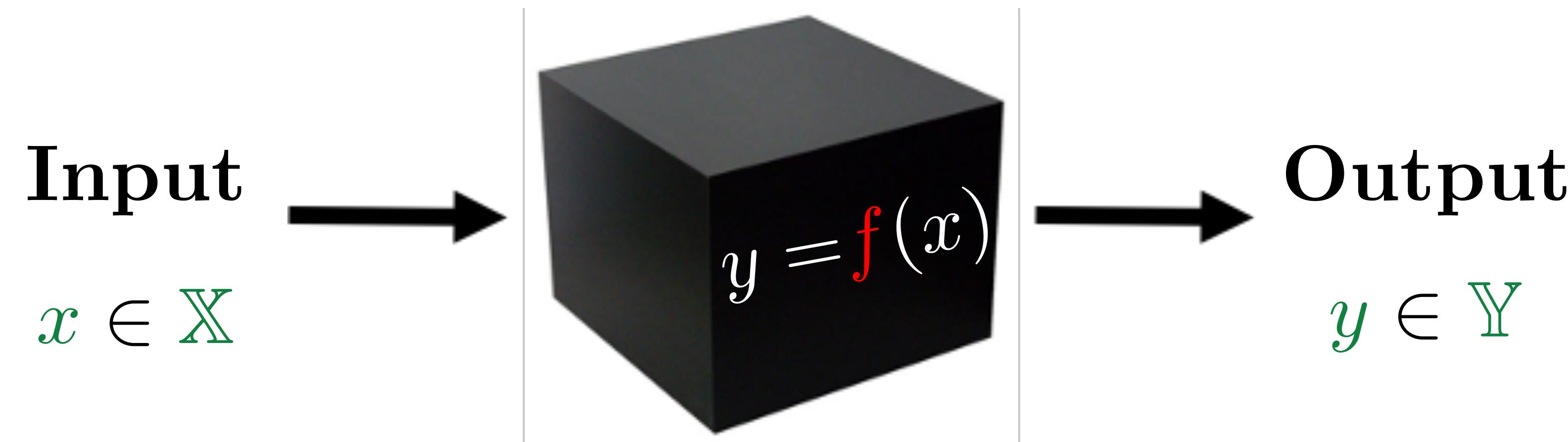
# Supervised Learning



# Supervised Learning

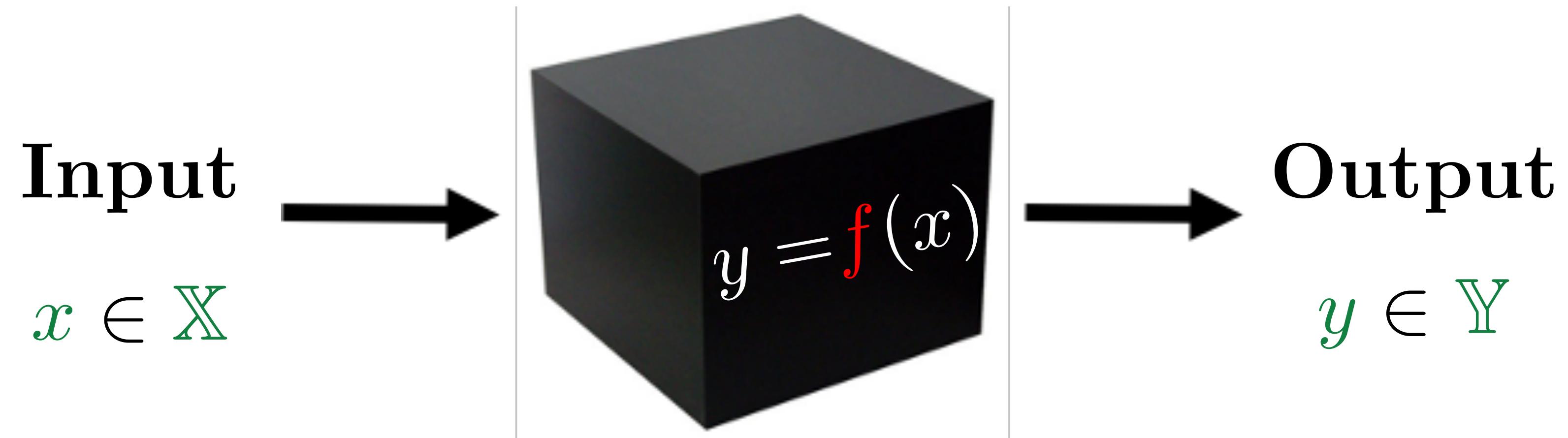


# Supervised Learning



Training data:  $\hat{\mathbb{E}}_N = (\hat{x}_1, \hat{y}_1), \dots, (\hat{x}_N, \hat{y}_N)$

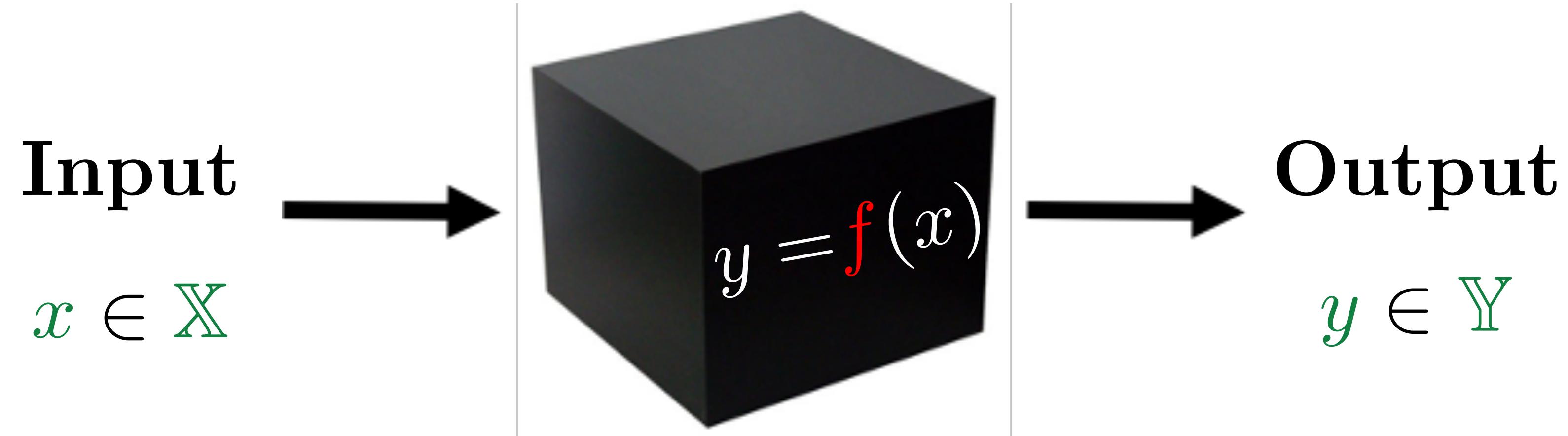
# Supervised Learning



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

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

# Supervised Learning

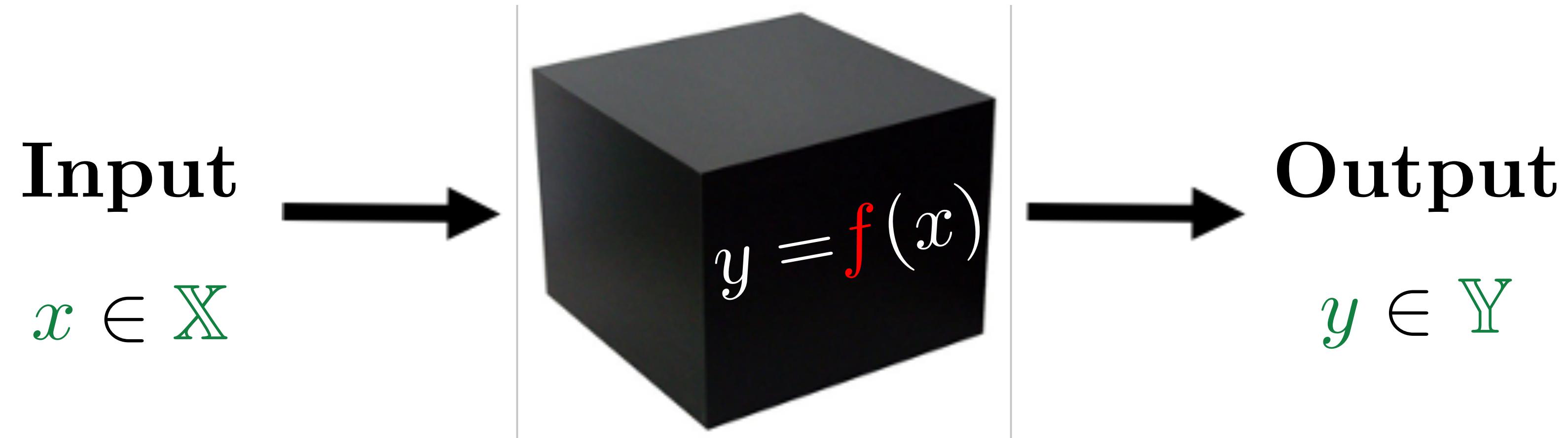


Training data:  $\hat{\Xi}_N = (\hat{x}_1, \hat{y}_1), \dots, (\hat{x}_N, \hat{y}_N)$

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

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

# Supervised Learning



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

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:  $f(\textcolor{violet}{x}) = \textcolor{red}{h}(\textcolor{violet}{x})$

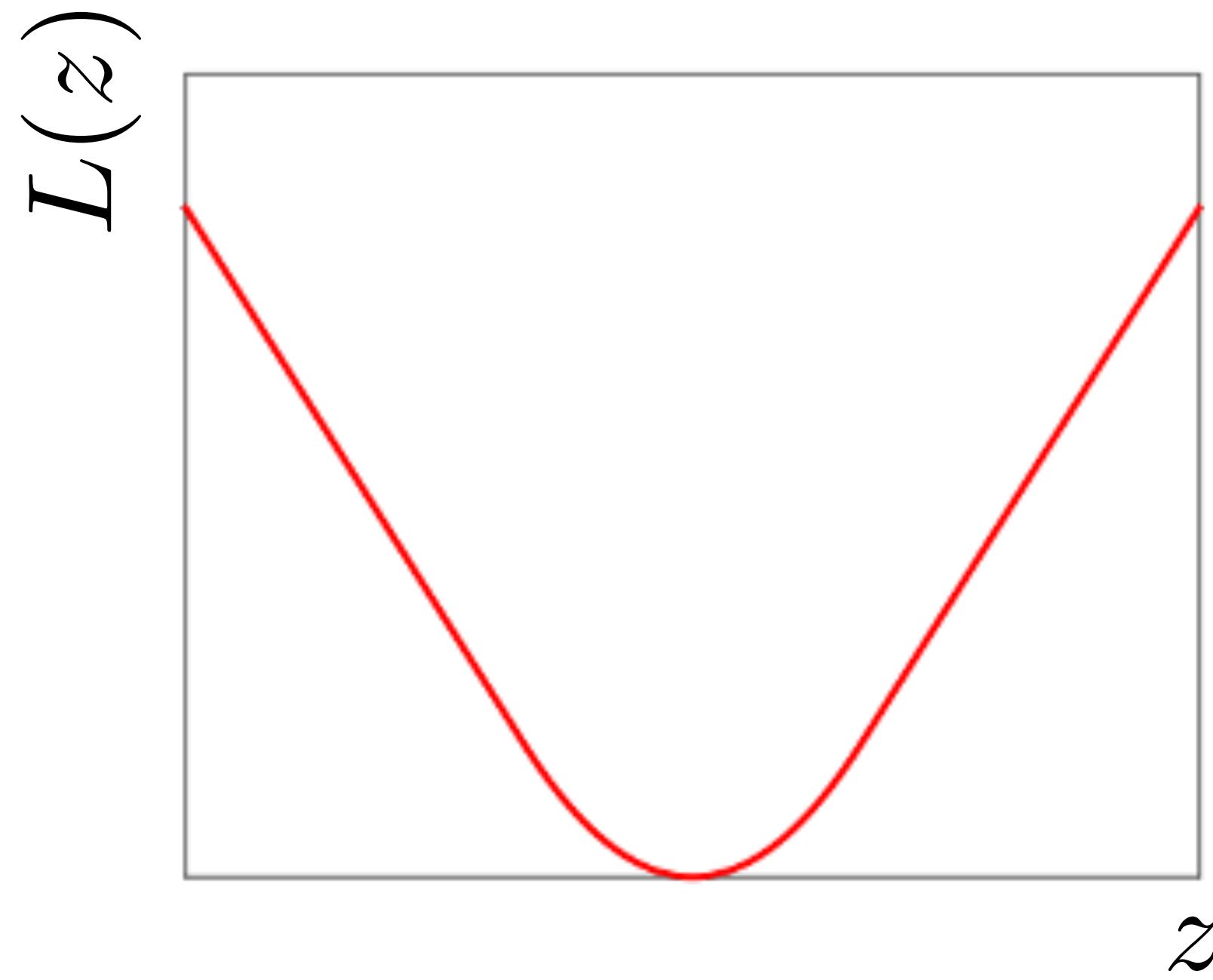
Empirical risk minimization:  $\ell(\textcolor{red}{h}, \widehat{\Xi}_N) = \frac{1}{N} \sum_{i=1}^N L(\textcolor{red}{h}(\widehat{x}_i) - \widehat{y}_i)$

# Regression Models

Target function:  $f(\textcolor{violet}{x}) = \textcolor{red}{h}(x)$

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

## Robust Regression



### Huber Loss:

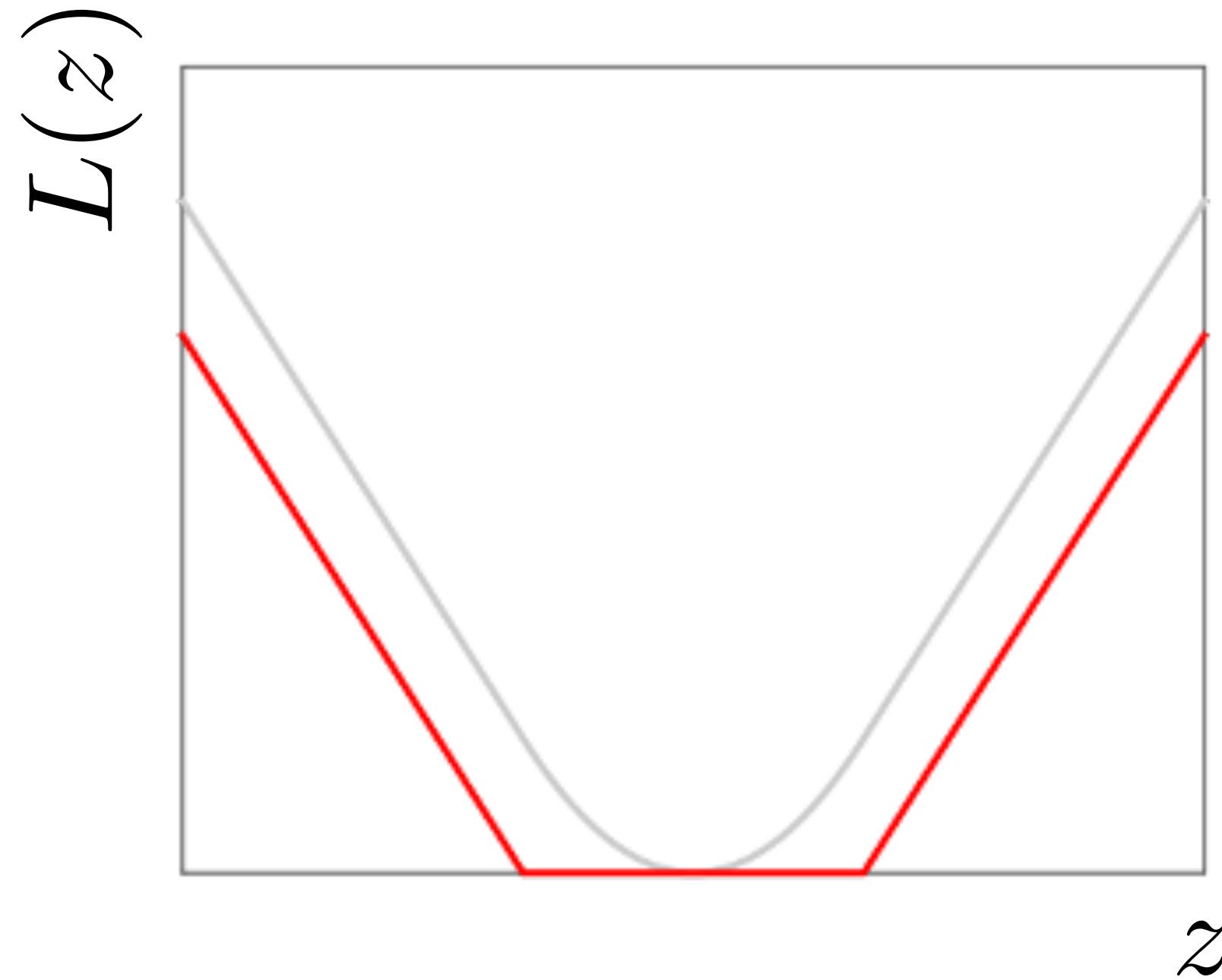
$$L(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta) & \text{else} \end{cases}$$

# Regression Models

Target function:  $f(\textcolor{violet}{x}) = \textcolor{red}{h}(\textcolor{violet}{x})$

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

## Support Vector Regression



$\epsilon$ -insensitive Loss:

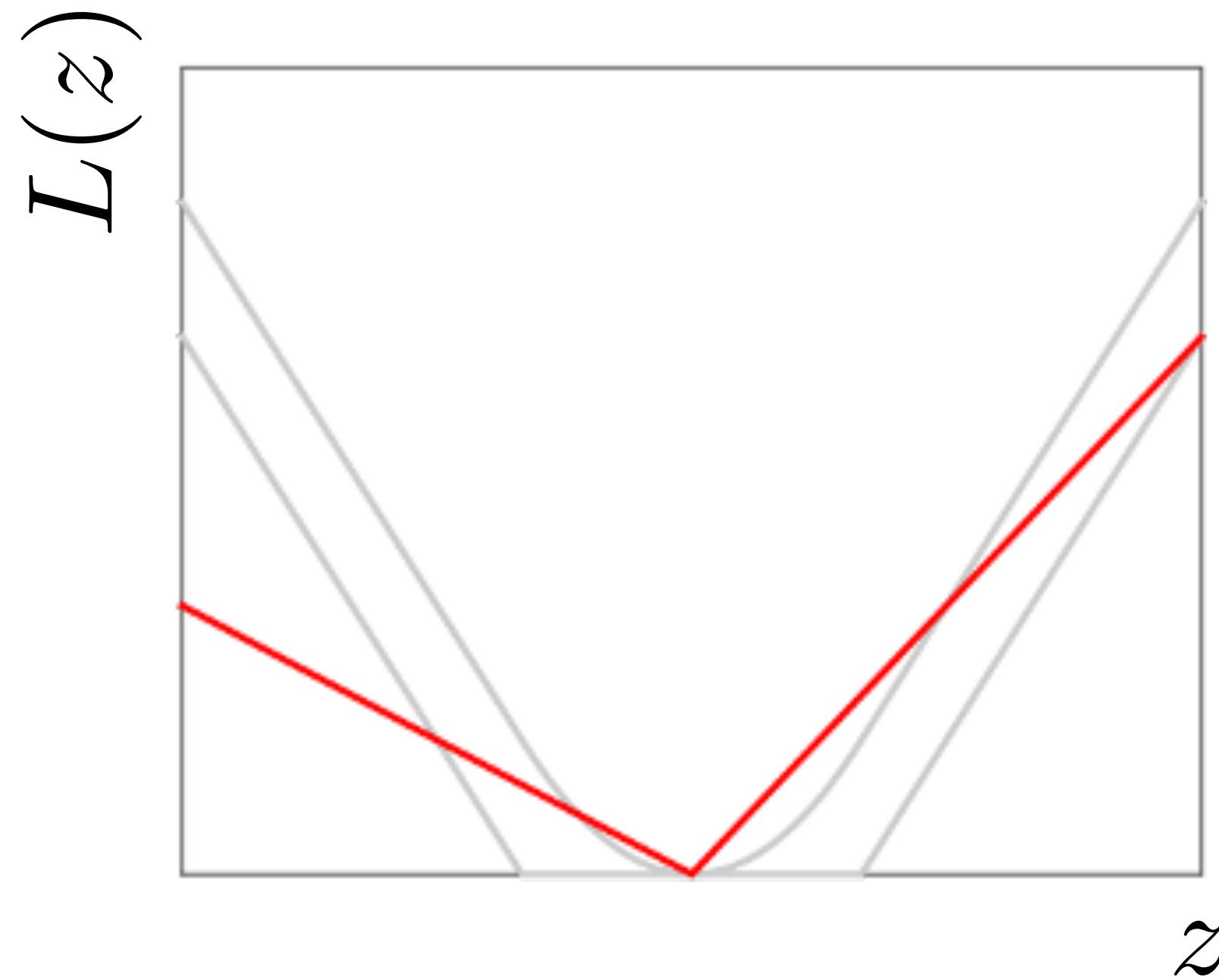
$$L(z) = \max\{0, |z| - \epsilon\}$$

# Regression Models

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

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

## Quantile Regression



### Pinball Loss:

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

# Classification Models

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

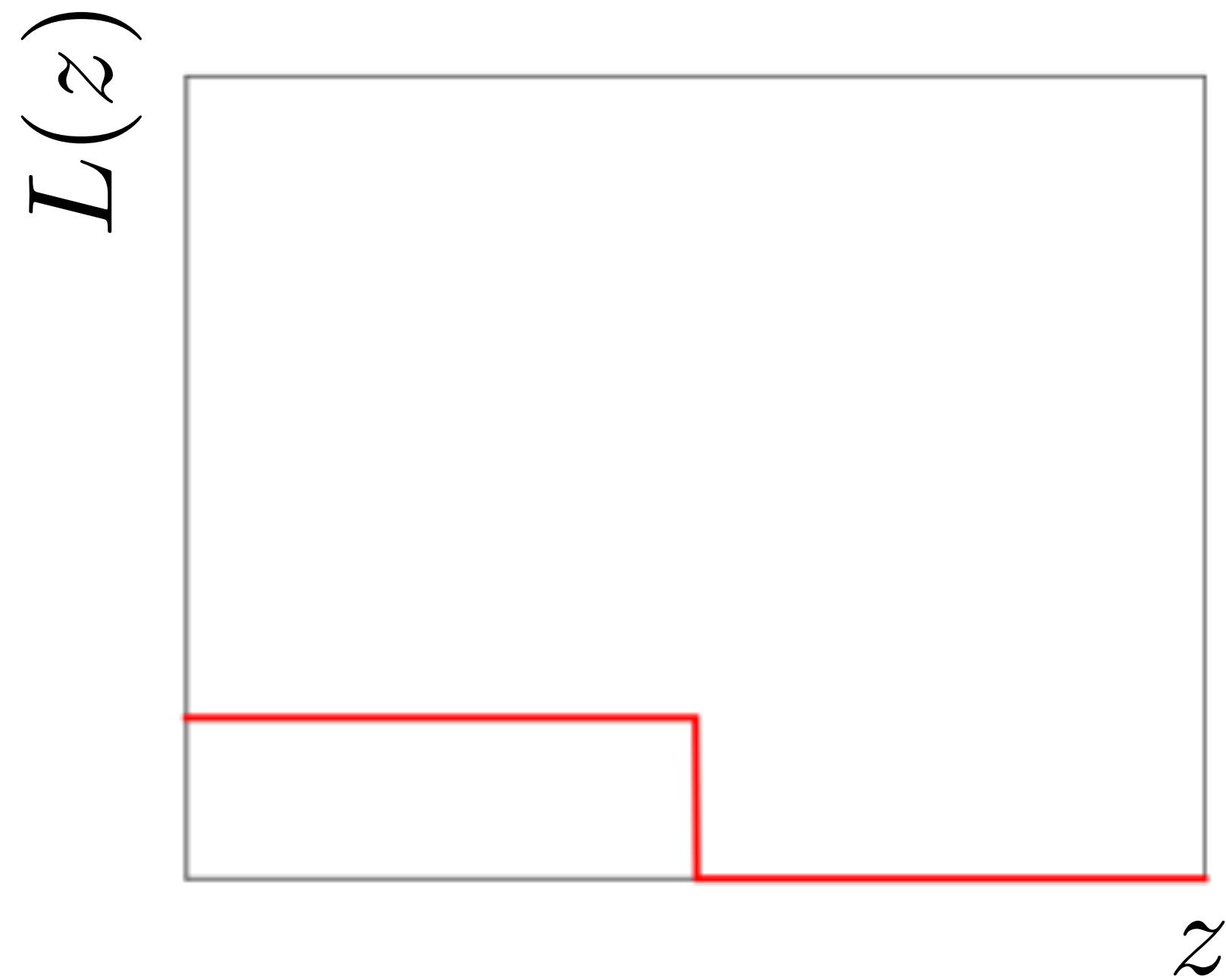
Empirical risk minimization:  $\ell(\textcolor{red}{h}, \widehat{\Xi}_{\textcolor{violet}{N}}) = \frac{1}{N} \sum_{i=1}^N L(\widehat{y}_i \textcolor{red}{h}(\widehat{x}_i))$

# Classification Models

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

Empirical risk minimization:  $\ell(\textcolor{red}{h}, \hat{\Sigma}_{\textcolor{teal}{N}}) = \frac{1}{N} \sum_{i=1}^N L(\hat{y}_i \textcolor{red}{h}(\hat{x}_i))$

## Ideal Classification



### 0-1 Loss:

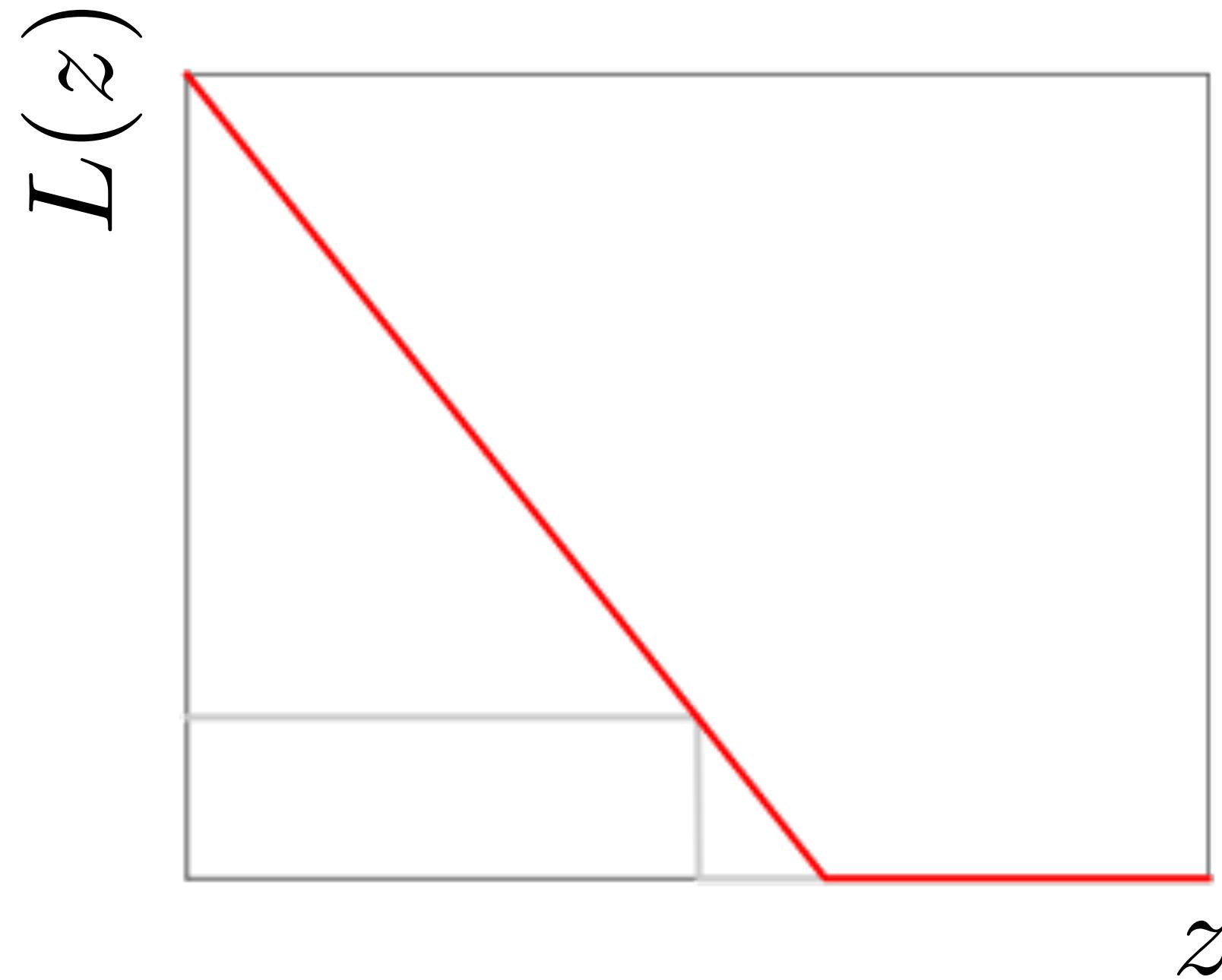
$$L(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 0 & \text{else} \end{cases}$$

# Classification Models

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

Empirical risk minimization:  $\ell(\textcolor{red}{h}, \hat{\Xi}_{\textcolor{teal}{N}}) = \frac{1}{N} \sum_{i=1}^N L(\hat{y}_i \textcolor{red}{h}(\hat{x}_i))$

## Support Vector Machine



### Hinge Loss:

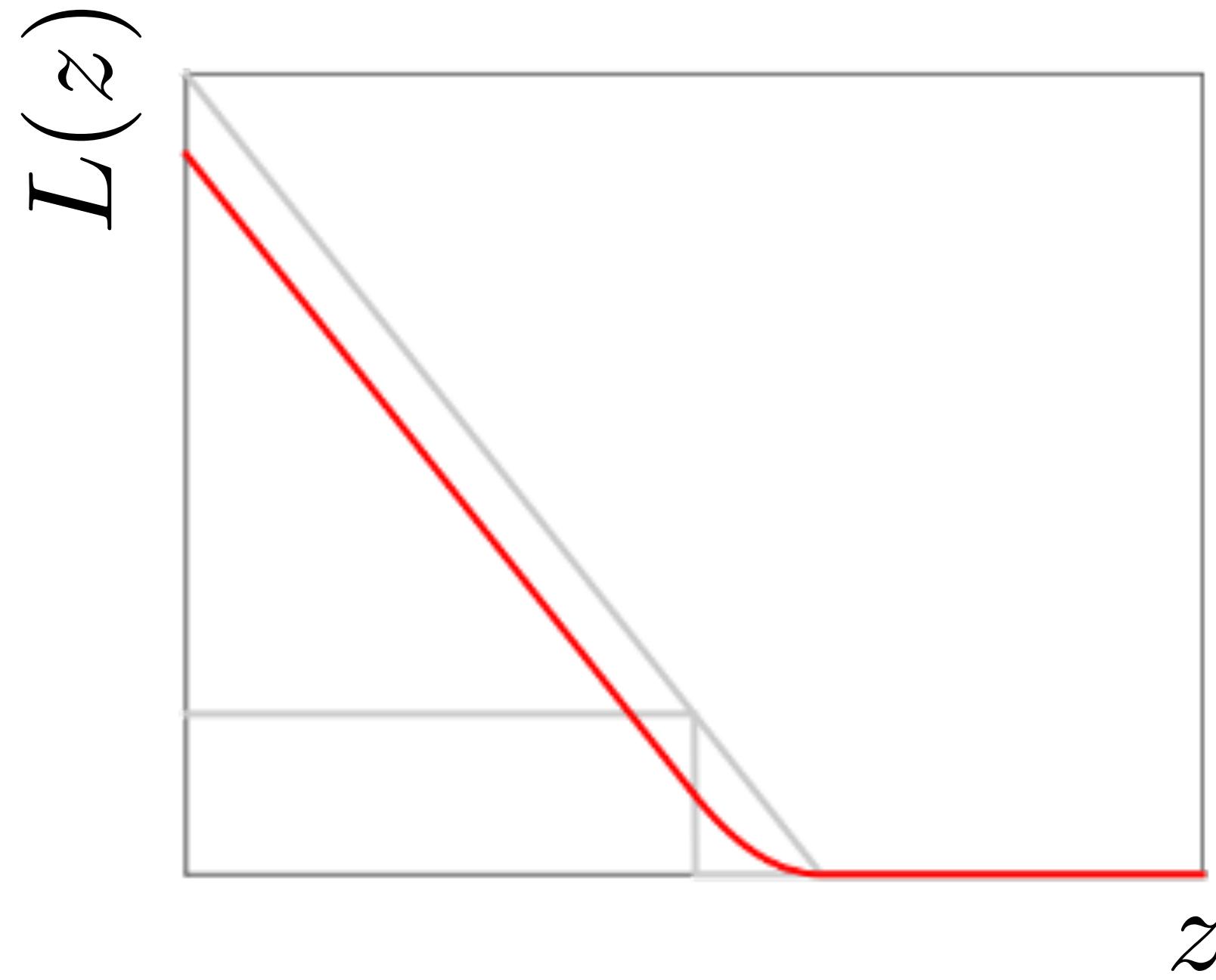
$$L(z) = \max\{0, 1 - z\}$$

# Classification Models

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

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

## Support Vector Machine II



Smooth Hinge Loss:

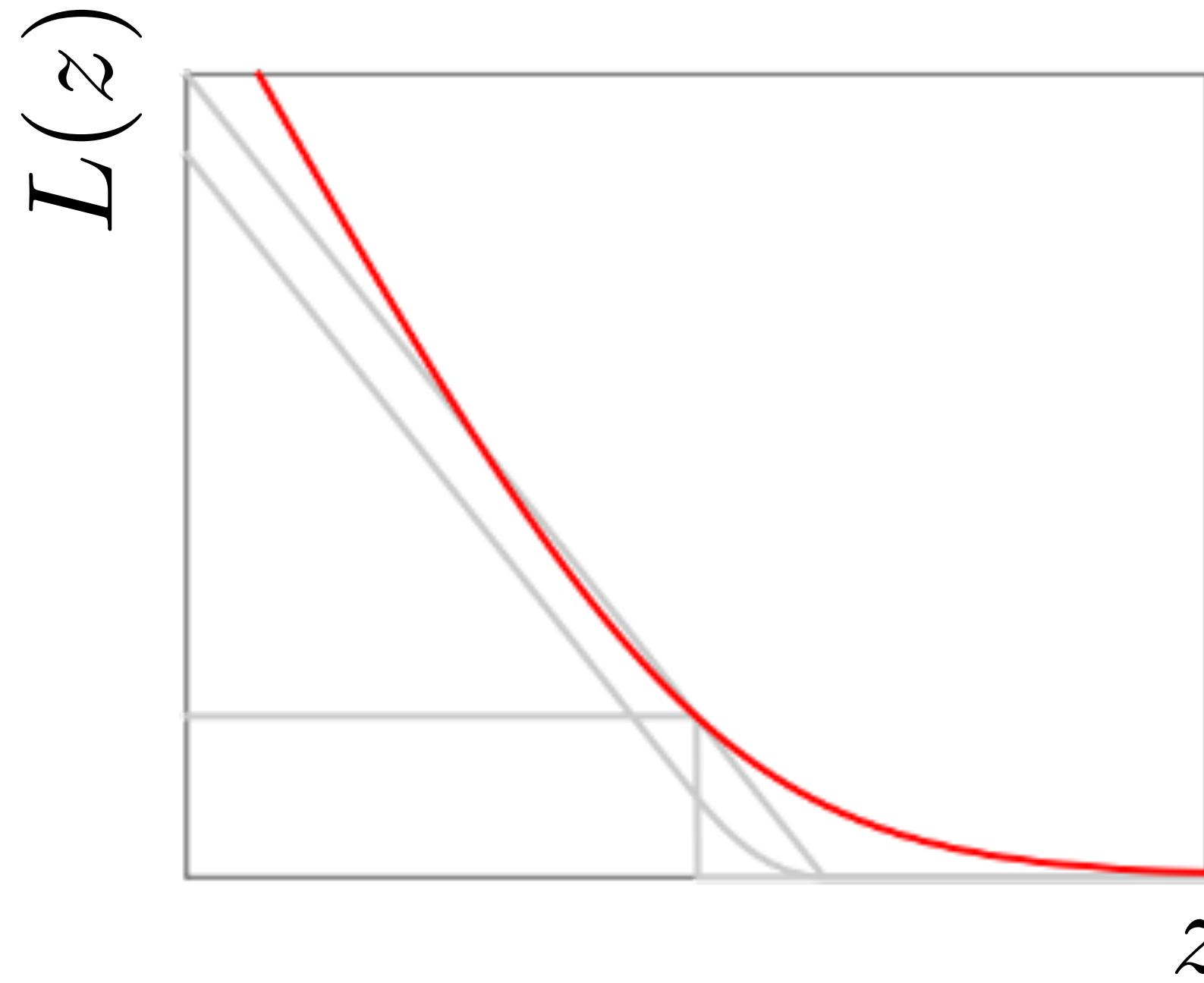
$$L(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1 - z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{else} \end{cases}$$

# Classification Models

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

Empirical risk minimization:  $\ell(\textcolor{red}{h}, \hat{\Sigma}_{\textcolor{teal}{N}}) = \frac{1}{N} \sum_{i=1}^N L(\hat{y}_i \textcolor{red}{h}(\hat{x}_i))$

## Logistic Regression



Logloss:

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

# Performance of ERM

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

In-Sample Performance:

$$\mathbb{E}_{\hat{\mathbb{P}}_N} [L(h(x), y)]$$

# Performance of ERM

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

In-Sample Performance:

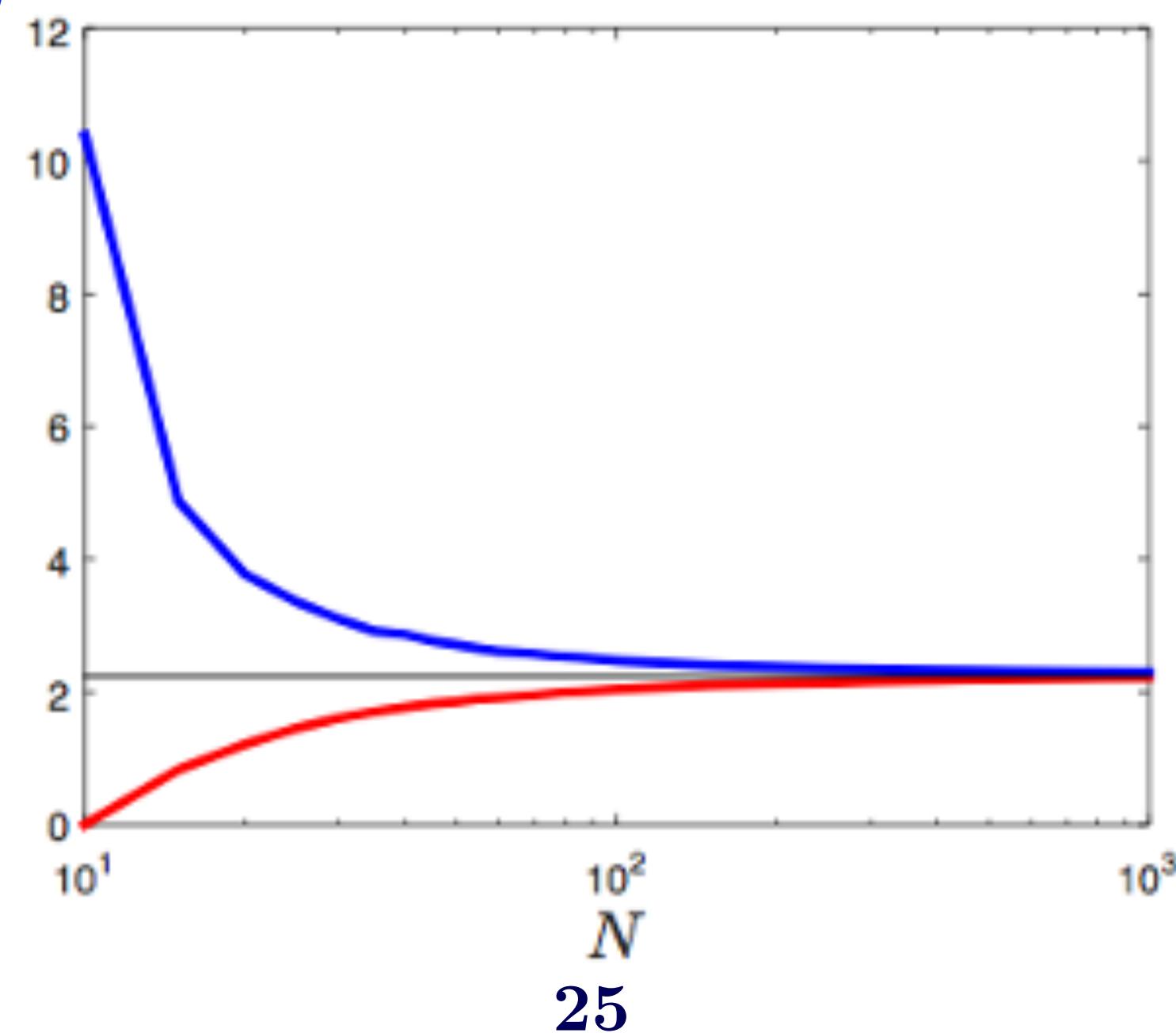
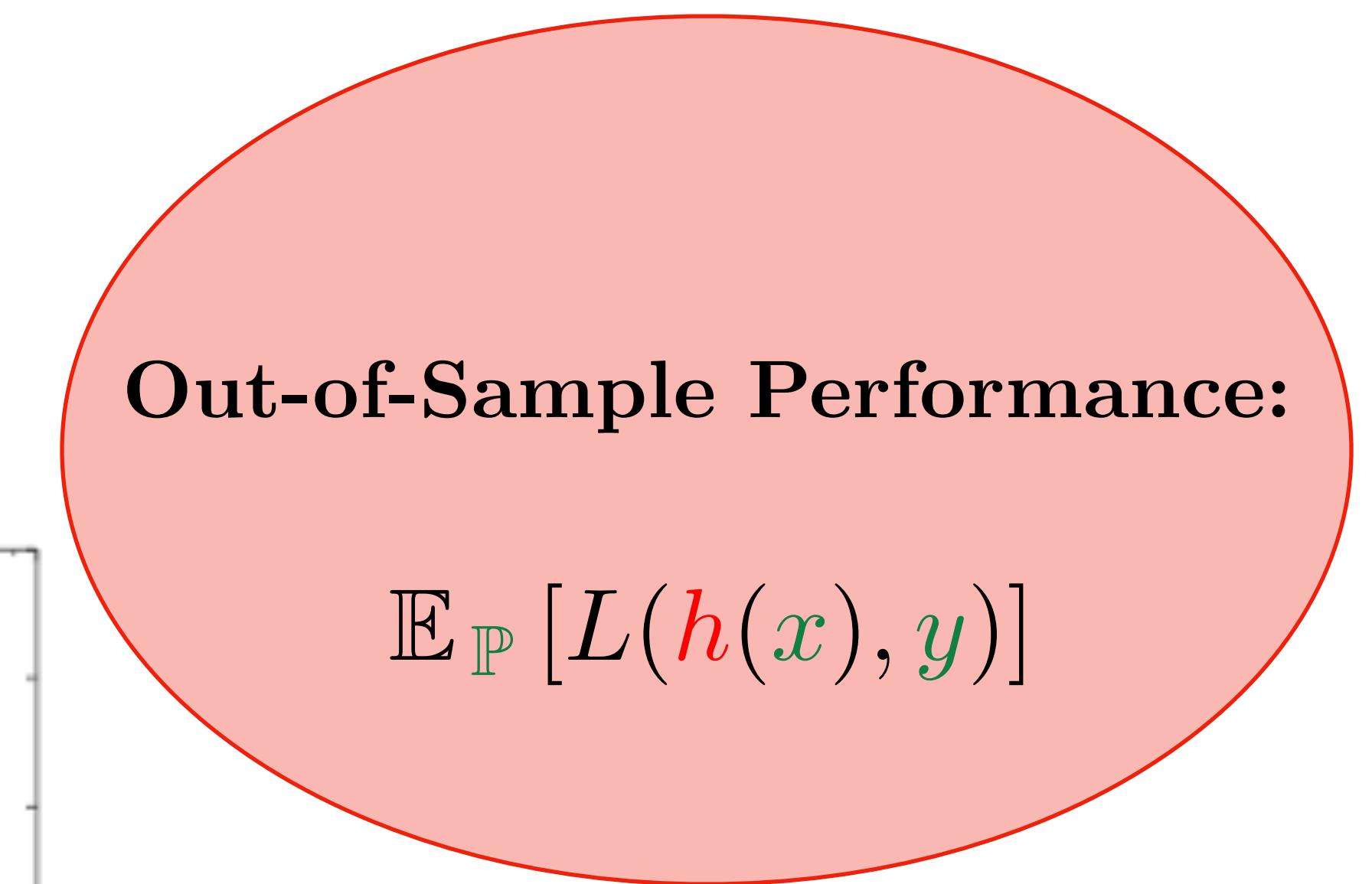
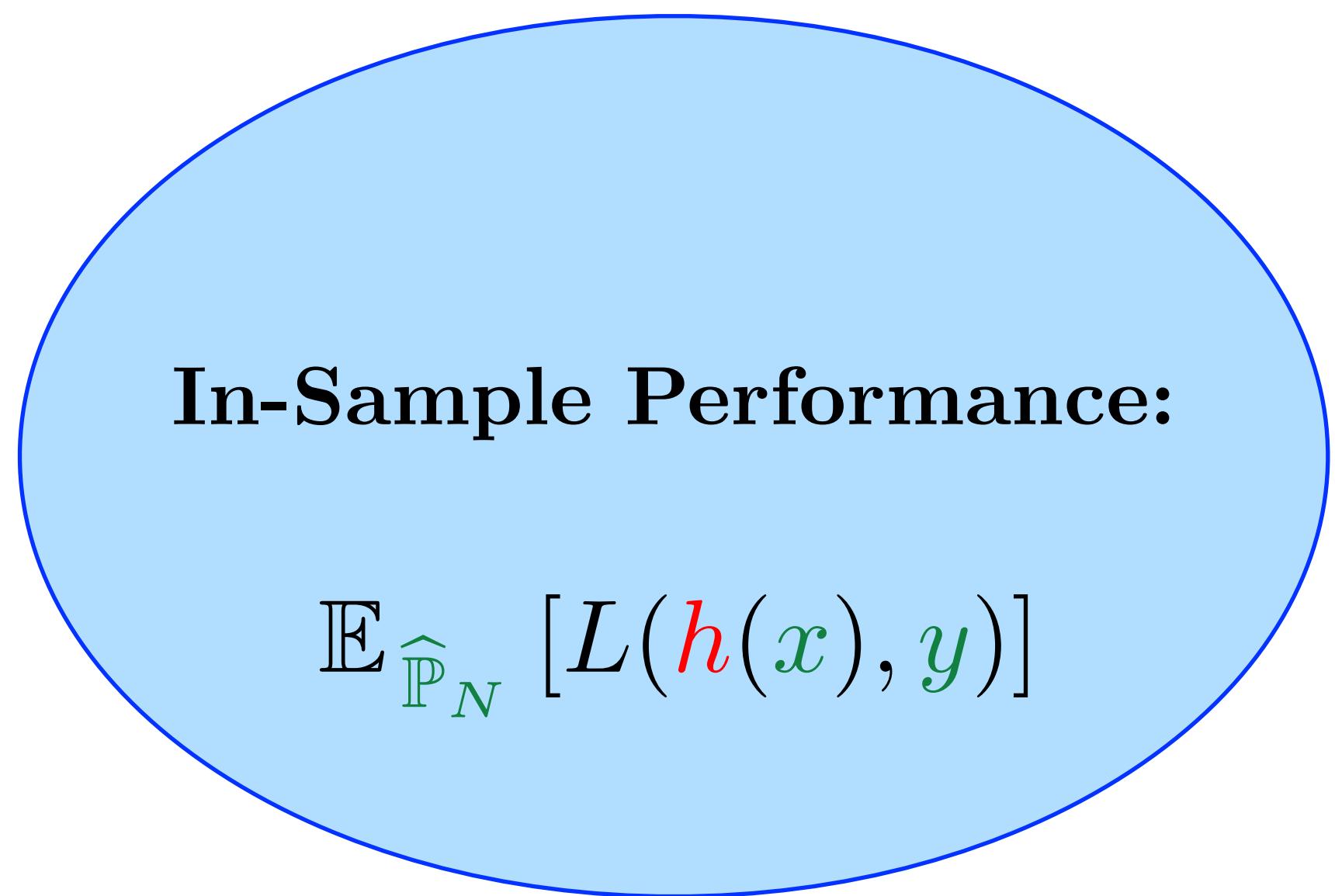
$$\mathbb{E}_{\hat{\mathbb{P}}_N} [L(h(x), y)]$$

Out-of-Sample Performance:

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

# Performance of ERM

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



# Regularized ERM

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

The diagram illustrates the components of the regularized ERM formula. A pink dotted circle highlights the regularization term  $\varepsilon \Omega(h)$ . A vertical pink arrow points from this circle down to the text 'Regularization coefficient'. A diagonal pink arrow points from the same circle to the text 'Regularization function'.

Regularization coefficient

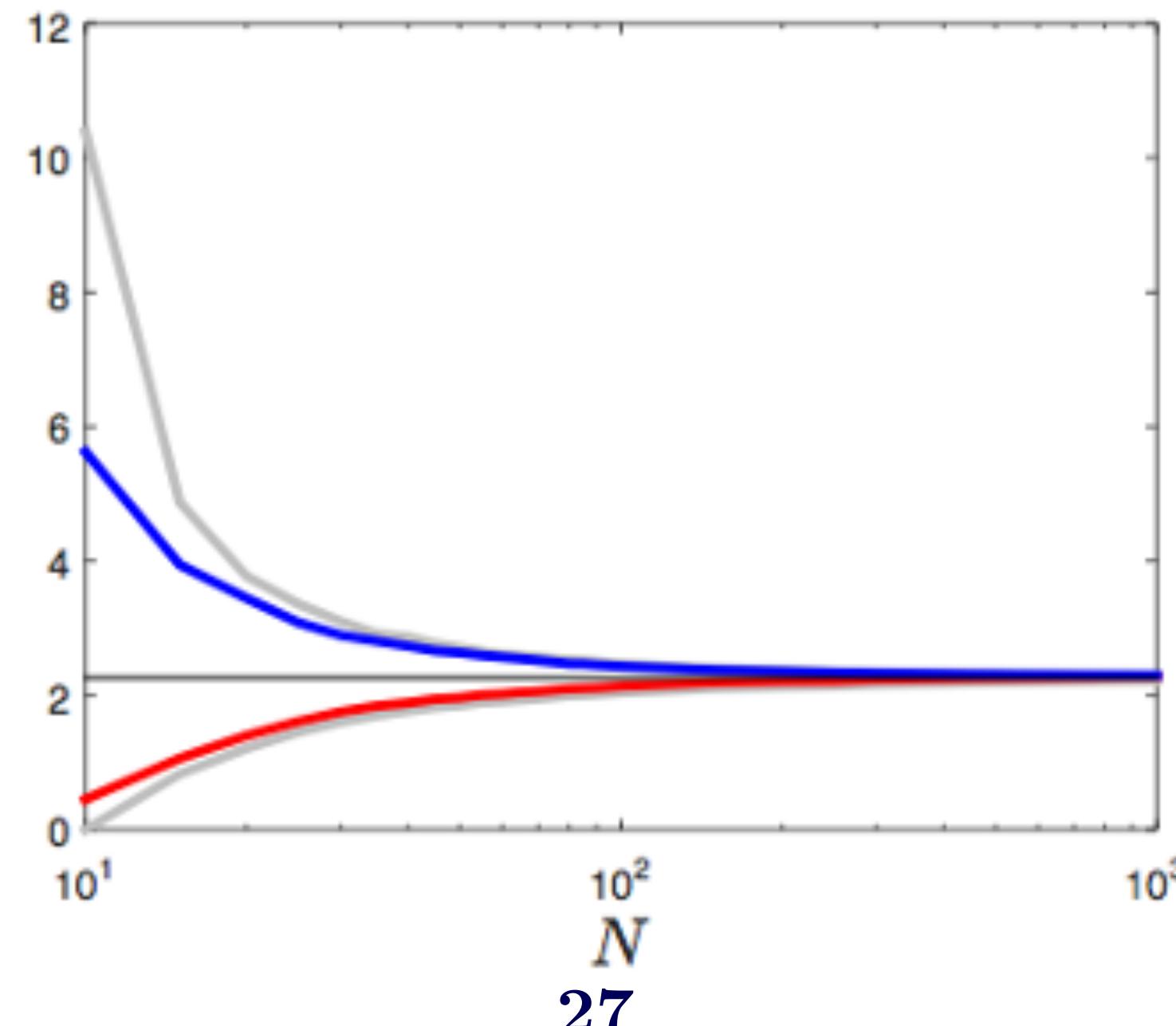
Regularization function

# Regularized ERM

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

Diagram illustrating the components of the regularized ERM formula:

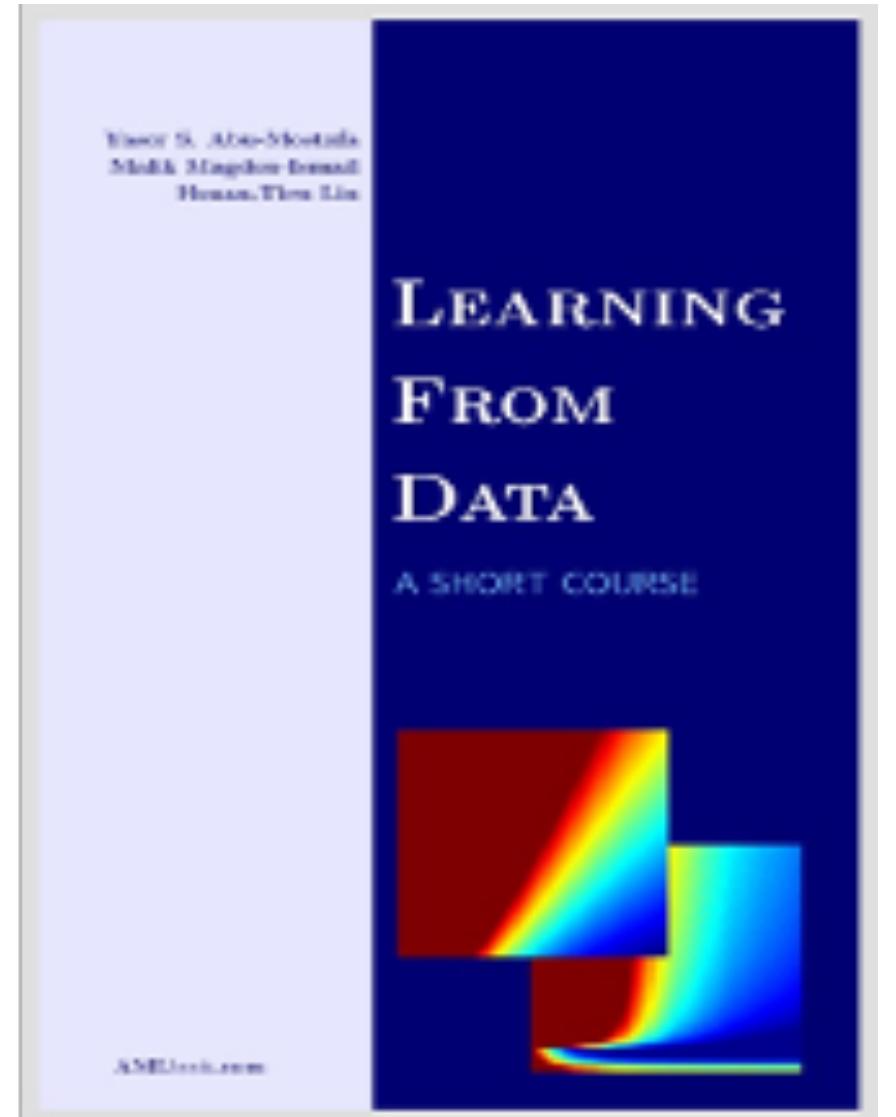
- Regularization coefficient**:  $\varepsilon$  (highlighted by a pink circle and arrow)
- Regularization function**:  $\Omega(h)$  (highlighted by a pink circle and arrow)



# Regularized ERM

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

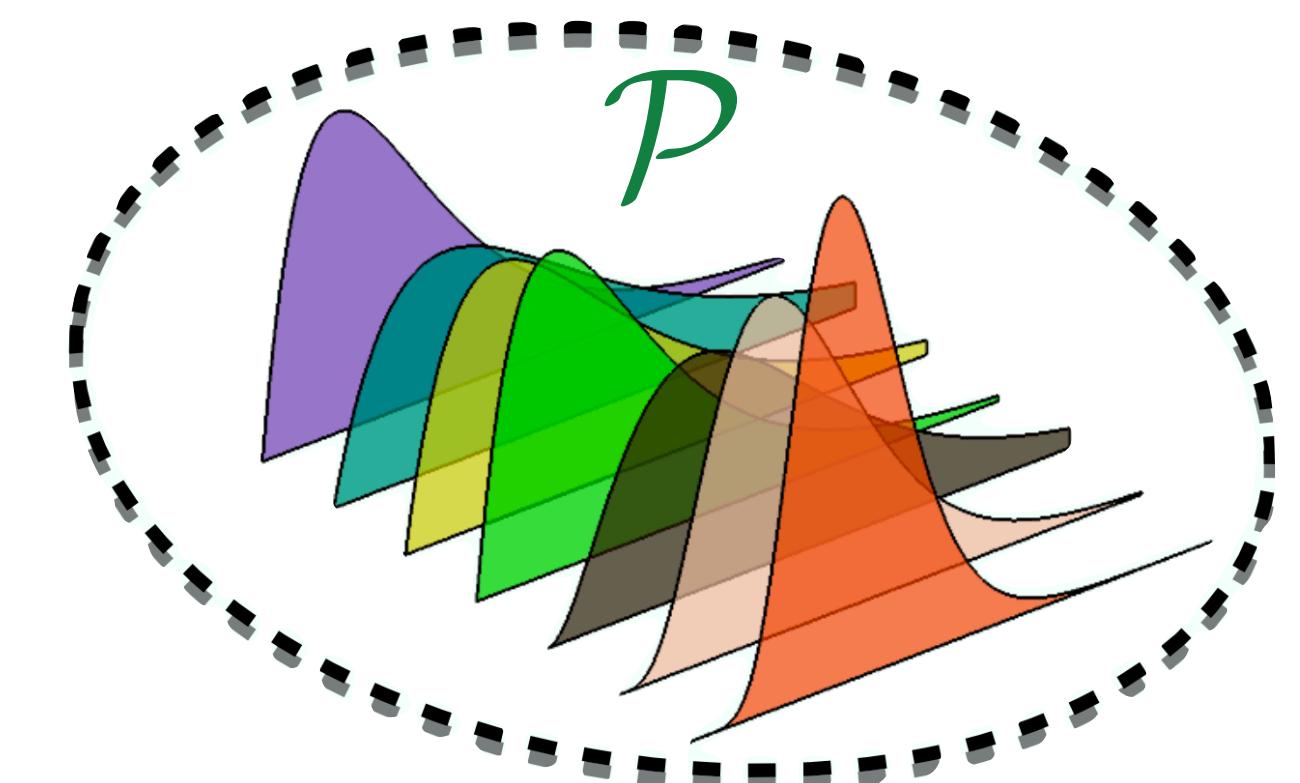
Abu-Mostafa *et al.*, 2012.



# Regularization via Optimal Transport

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

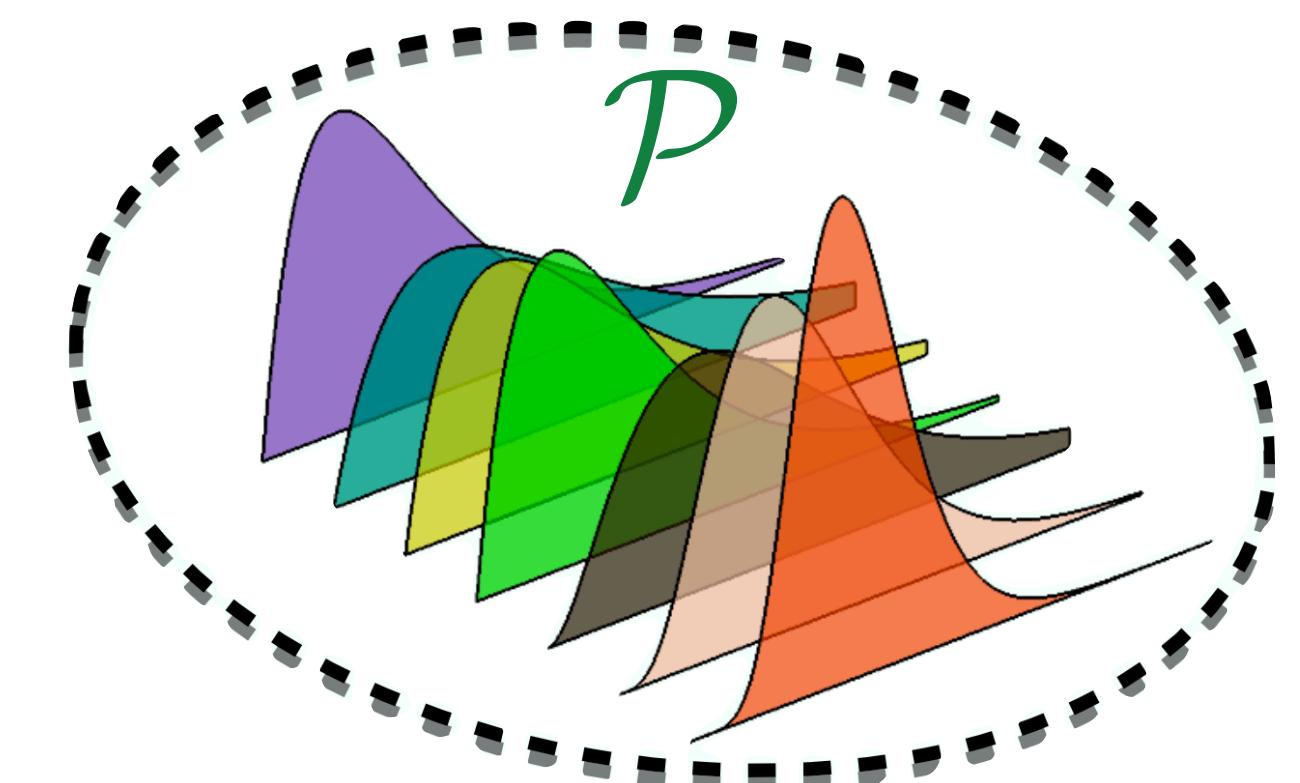
$$\mathbb{H} = \{ \color{red} h \in \mathbb{R}^X : \exists \theta \in \Theta \text{ s.t. } \color{red} h(x) = \theta^\top x \}$$



# Regularization via Optimal Transport

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

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



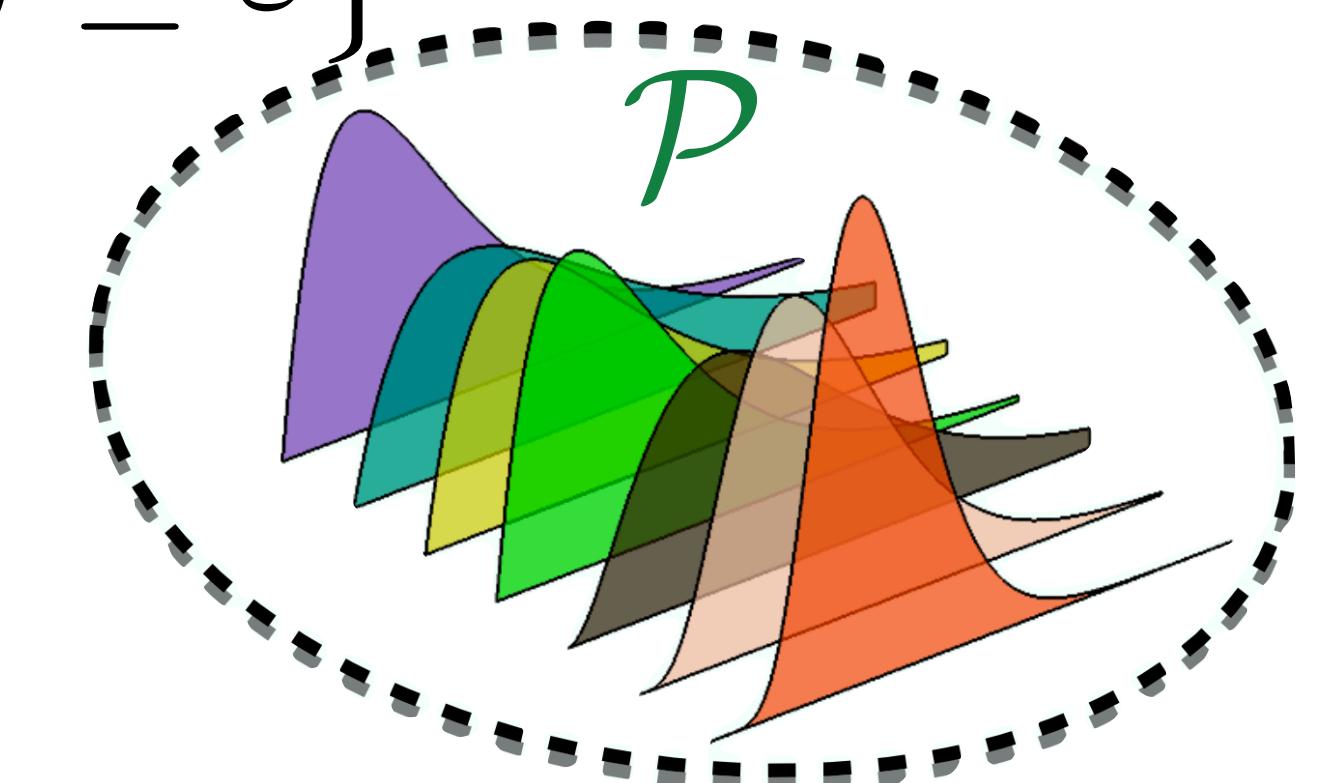
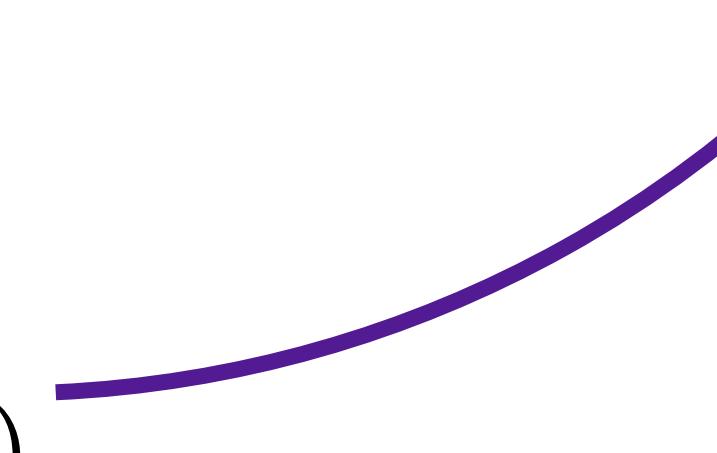
# Regularization via Optimal Transport

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

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

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

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



# The Real Story Behind the Success

## Regularization

[SMK15, GCK17, CP18,  
BMZ18, BKM19, SKM19]

## Statistical Guarantees

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

# Optimal Transport: Old and New



Monge

Hitchcock

Kantorovich-Koopmans

Vaserstein

Brenier

Villani

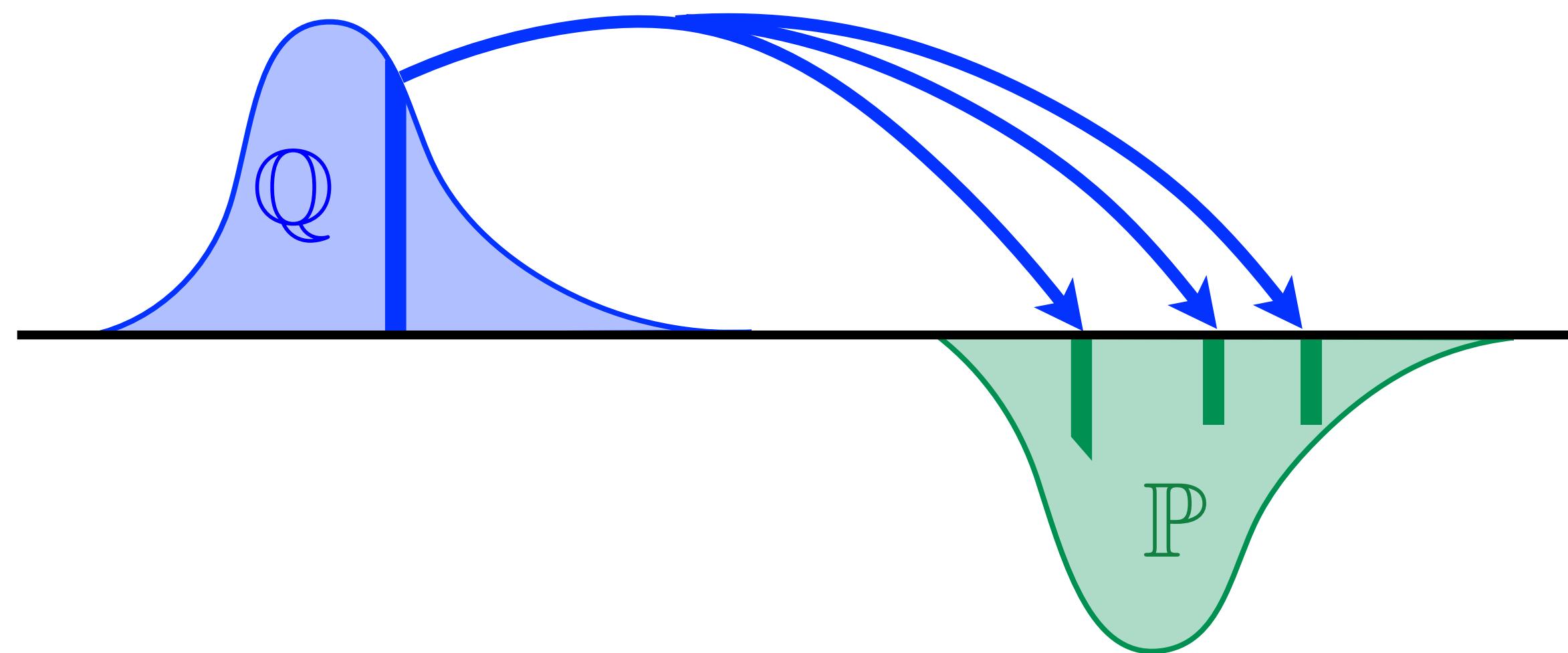
Figalli

Nobel '75

Fields '10

Fields '18

# What is Optimal Transport?



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



THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES

BY FRANK L. HITCHCOCK

1941

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

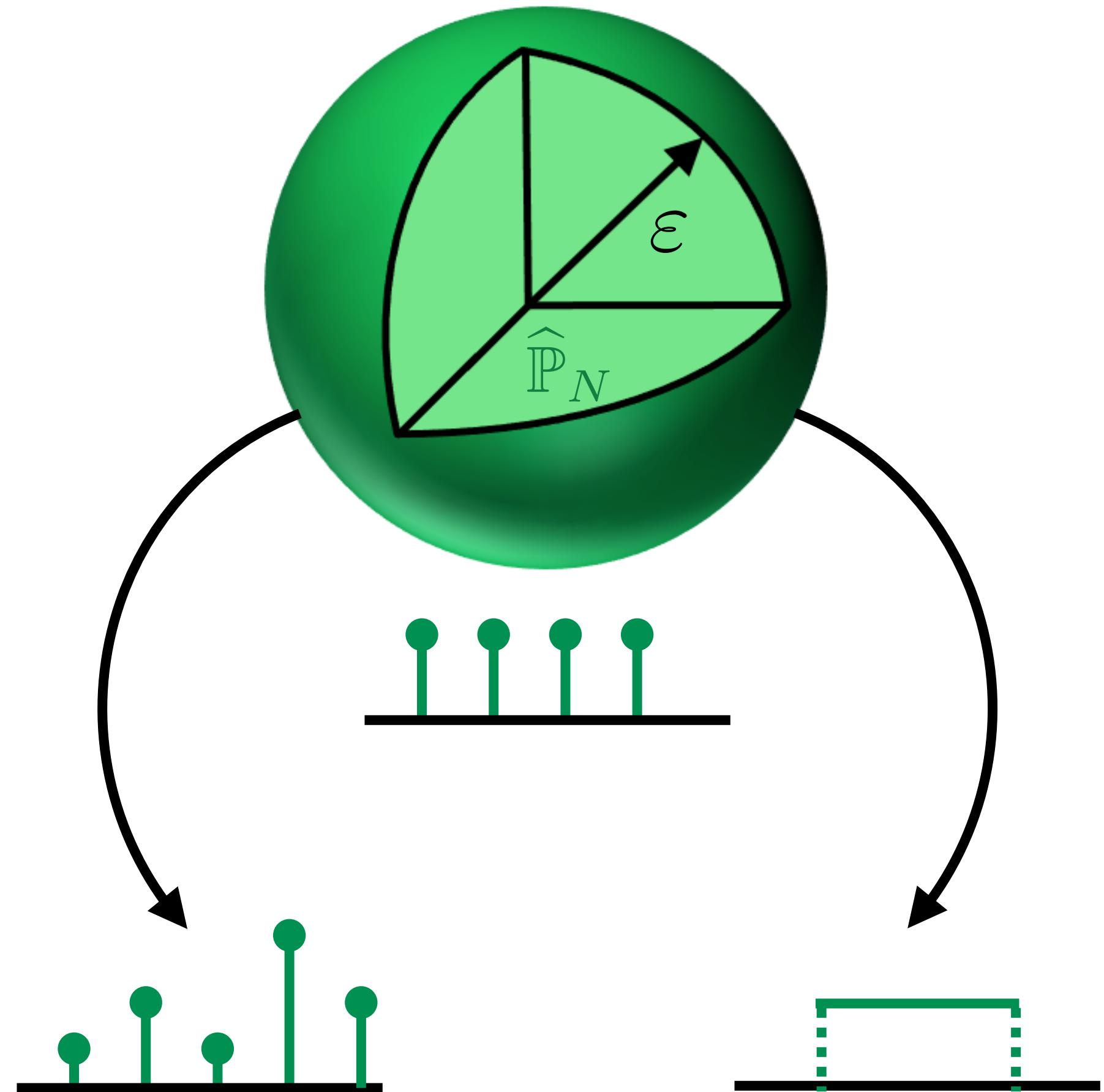
by Tjalling C. Koopmans

*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

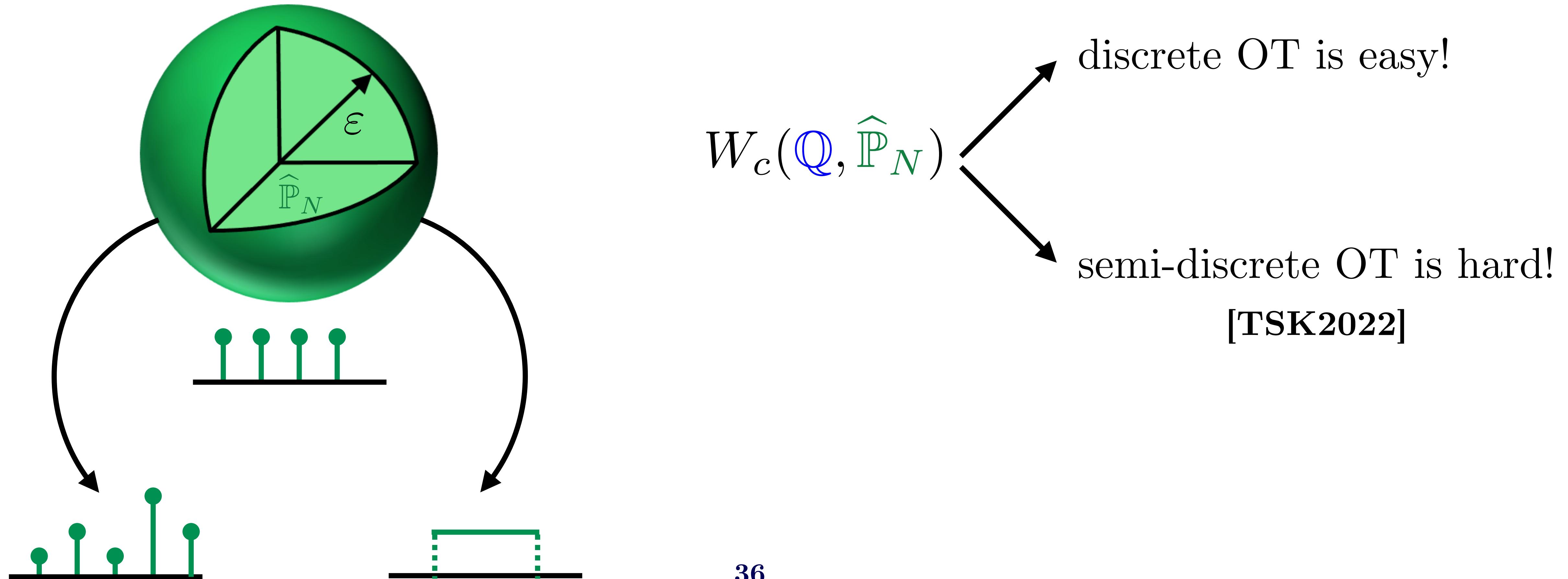
# Optimal Transport Ambiguity Set

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



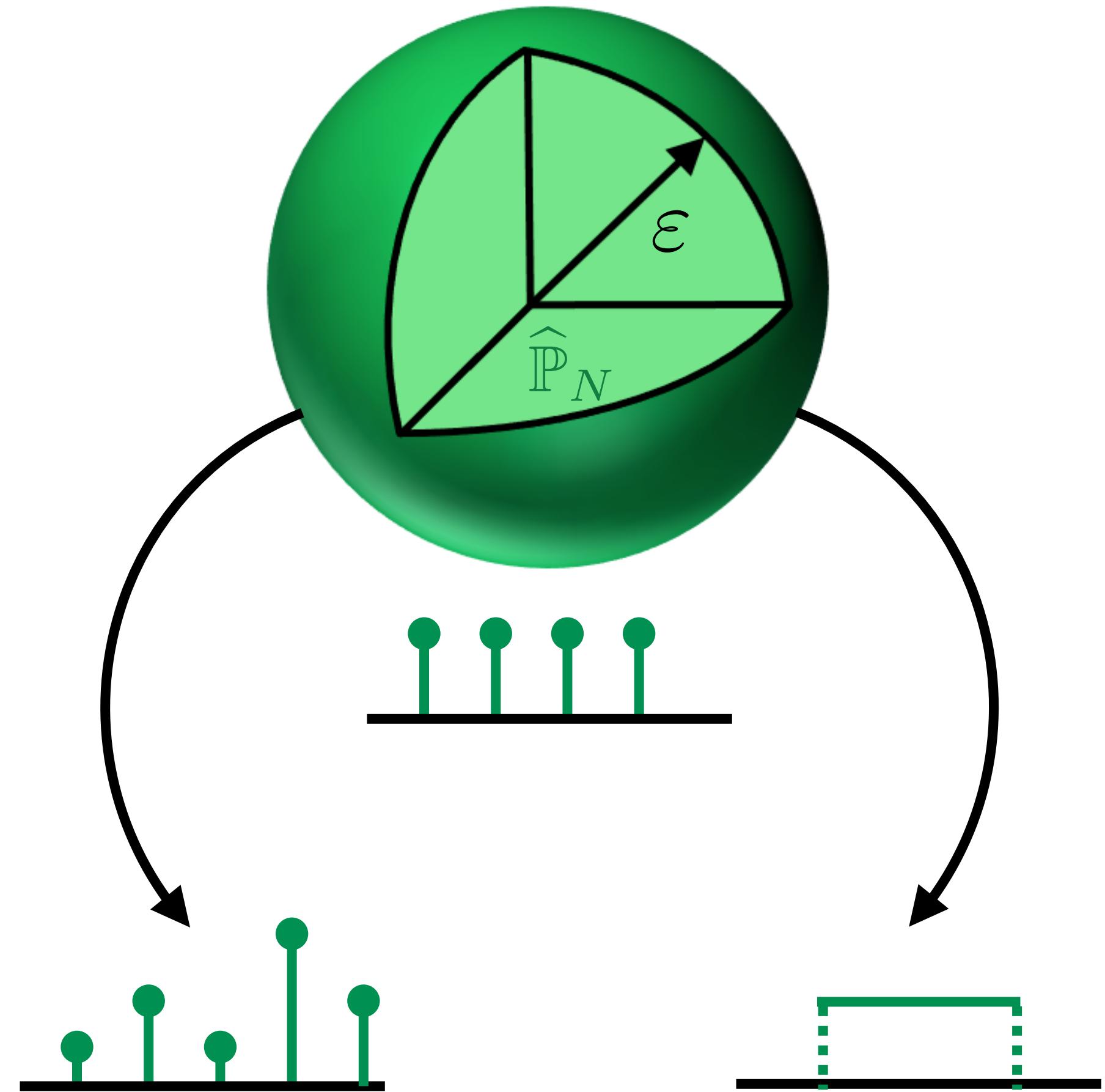
# Optimal Transport Ambiguity Set

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



# Optimal Transport Ambiguity Set

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



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



# Tractability for Linear Regression

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

If  $L$  is convex and Lipschitz, then

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

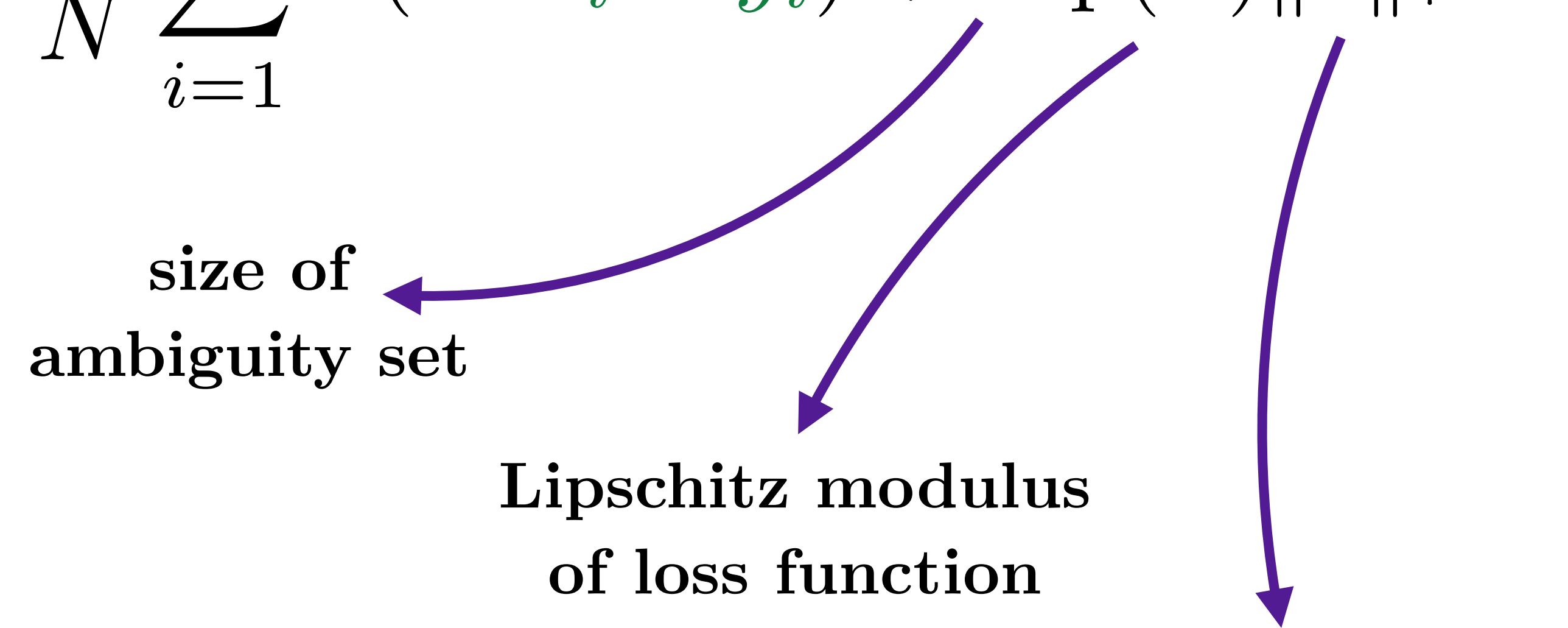
[SMK15, GCK17, SKM19, BKM19]

# Tractability for Linear Regression

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

If  $L$  is convex and Lipschitz, then

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



[SMK15, GCK17, SKM19, BKM19]

# Semi-infinite Duality

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

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

[MK18, ZG18, BM19, GK16, ZYG22]

# Proof of Lemma 1

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

# Proof of Lemma 1

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

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

# Proof of Lemma 1

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

↓

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

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

# Proof of Lemma 1

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

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

# Proof of Lemma 1

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} [I(\xi)] = \left\{ \begin{array}{l} \sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} \frac{1}{N} \sum_{i=1}^N \int_{\xi \in \Xi} I(\xi) \mathbb{Q}_i(d\xi) \\ \frac{1}{N} \int_{\xi \in \Xi} c(\xi, \hat{\xi}_i) \mathbb{Q}_i(d\xi) \leq \varepsilon \end{array} \right.$$

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

$$= \left\{ \begin{array}{l} \sup_{\mathbb{Q}_i \geq 0} \frac{1}{N} \sum_{i=1}^N \int_{\xi \in \Xi} I(\xi) \mathbb{Q}_i(d\xi) \\ \frac{1}{N} \int_{\xi \in \Xi} c(\xi, \hat{\xi}_i) \mathbb{Q}_i(d\xi) \leq \varepsilon \\ \int_{\xi \in \Xi} \mathbb{Q}_i(d\xi) = 1 \quad \forall i \in [N] \end{array} \right.$$

# Proof of Lemma 1

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}} [I(\xi)] = \left\{ \begin{array}{l} \sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi)} \frac{1}{N} \sum_{i=1}^N \int_{\xi \in \Xi} I(\xi) \mathbb{Q}_i(d\xi) \\ \frac{1}{N} \int_{\xi \in \Xi} c(\xi, \hat{\xi}_i) \mathbb{Q}_i(d\xi) \leq \varepsilon \end{array} \right.$$

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

$$= \left\{ \begin{array}{l} \sup_{\mathbb{Q}_i \geq 0} \frac{1}{N} \sum_{i=1}^N \int_{\xi \in \Xi} I(\xi) \mathbb{Q}_i(d\xi) \\ \frac{1}{N} \int_{\xi \in \Xi} c(\xi, \hat{\xi}_i) \mathbb{Q}_i(d\xi) \leq \varepsilon \quad (\lambda) \\ \int_{\xi \in \Xi} \mathbb{Q}_i(d\xi) = 1 \quad (s_i) \quad \forall i \in [N] \end{array} \right.$$

# Proof of Lemma 1

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

# Proof of Lemma 1

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

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

# Proof of Lemma 1

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

$$\begin{aligned} &= \inf_{\lambda \geq 0} \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^N \sup_{\xi \in \Xi} I(\xi) - \lambda c(\xi, \widehat{\xi}_i) \\ &= \inf_{\lambda \geq 0} \lambda \varepsilon + \mathbb{E}_{\widehat{\mathbb{P}}_N} \left[ \sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi) \right] \end{aligned}$$

# Proof of Lemma 1

$$\sup_{Q \in \mathcal{P}} \mathbb{E}_Q [I(\xi)] \stackrel{\text{[S01]}}{\leq} \left\{ \begin{array}{ll} \inf_{\substack{\lambda \geq 0 \\ s \in \mathbb{R}^N}} & \lambda \varepsilon + \sum_{i=1}^N s_i \\ \text{s.t.} & \frac{\lambda}{N} c(\xi, \hat{\xi}_i) + s_i \geq \frac{1}{N} I(\xi) \quad \forall \xi \in \Xi \end{array} \right.$$

$$\begin{aligned} &= \inf_{\lambda \geq 0} \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^N \sup_{\xi \in \Xi} I(\xi) - \lambda c(\xi, \hat{\xi}_i) \\ &= \inf_{\lambda \geq 0} \lambda \varepsilon + \mathbb{E}_{\hat{\mathbb{P}}_N} \left[ \sup_{\zeta \in \Xi} I(\zeta) - \lambda c(\zeta, \xi) \right] \end{aligned}$$

# 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

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

# Lipschitz Envelope

**Lemma 2.** Let  $L : \mathbb{R} \rightarrow \mathbb{R}$  be a convex and Lipschitz function.

Then,

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

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

# 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

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

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

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

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

# Proof of Lemma 2

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

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

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

# Proof of Lemma 2

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

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

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

## 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

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

# Proof of Lemma 2

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

# Proof of Lemma 2

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

# Tractability for Linear Regression

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

If  $L$  is convex and Lipschitz, then

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

[SMK15, GCK17, SKM19, BKM19]

# Proof of Theorem 1

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

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

# 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

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

[MK18, ZG18, BM19, GK16, ZYG22]

# Proof of Theorem 1

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

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

# Proof of Theorem 1

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

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

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

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

# Lipschitz Envelope

**Lemma 2.** Let  $L : \mathbb{R} \rightarrow \mathbb{R}$  be a convex and Lipschitz function.

Then,

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

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

# Proof of Theorem 1

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

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

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

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

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

# Tractability for Linear Regression

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

If  $L$  is convex and Lipschitz, then

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

# Examples

## Robust Regression

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

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



## 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

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

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



# Least Squares Regression

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

If  $L = z^2$ , then

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

[BKM19]

# Moreau Envelope

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

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

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

[BKM19, SADK22]

# Proof of Lemma 3

$$\begin{aligned} & \sup_{\zeta} L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 \\ &= \sup_{\Delta} L(\theta^\top \Delta + \theta^\top \xi + \theta_0) - \lambda \|\Delta\|^2 \end{aligned}$$

# Proof of Lemma 3

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

# Proof of Lemma 3

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

# Proof of Lemma 3

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

# Proof of Lemma 3

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

# 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

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

# Proof of Lemma 3

$$\begin{aligned} \sup_{\zeta} L(\theta^\top \zeta + \theta_0) - \lambda \|\zeta - \xi\|^2 &= \sup_{\gamma} L(\gamma + \theta^\top \xi) - \frac{\lambda \gamma^2}{\|\theta\|_*^2} \\ &= \sup_{\gamma} (\gamma + \theta^\top \xi)^2 - \frac{\lambda \gamma^2}{\|\theta\|_*^2} \\ &= \begin{cases} \frac{\lambda}{\lambda - \|\theta\|_*^2} (\theta^\top \xi)^2 & \text{if } \|\theta\|_*^2 < \lambda \\ +\infty & \text{else} \end{cases} \end{aligned}$$

# Least Squares Regression

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

If  $L = z^2$ , then

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

[BKM19]

# Proof of Theorem 2

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

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

# 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

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

[MK18, ZG18, BM19, GK16, ZYG22]

# Proof of Theorem 2

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

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

# Proof of Theorem 2

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

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

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

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

# Proof of Theorem 2

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

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

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

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

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

# Tractability for Linear Classification

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

If  $L$  is convex and Lipschitz, then

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

[SMK15, GCK17, SKM19, BKM19]

# Proof of Theorem 3

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

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

# 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 Theorem 3

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

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

# Proof of Theorem 3

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

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

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

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

# Proof of Theorem 3

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

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

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

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

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

# Examples

## Support Vector Machine

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

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



## Support Vector Machine II

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

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



## Logistic Regression

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

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

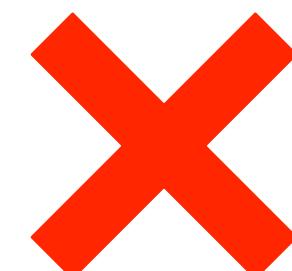


# Examples

Ideal  
Classification

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

nonconvex!



# Ideal Classification

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

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

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

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

# Ideal Classification

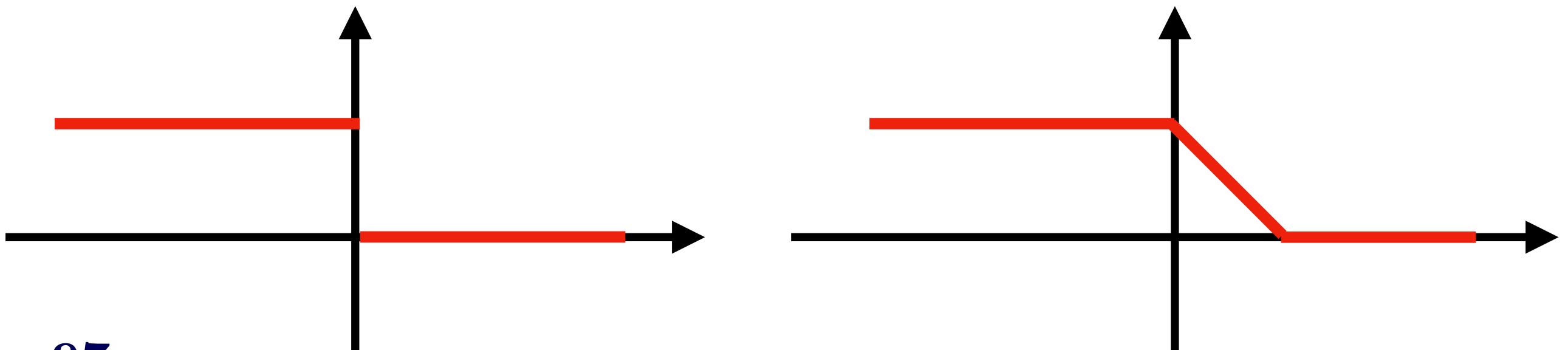
**Theorem 4.** Suppose that  $c((x, y), (x', y')) = \|x - x'\| + \delta_{y=y'}.$

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

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

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,

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

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

# 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

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

# Proof of Lemma 4

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

# Proof of Lemma 4

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

# Proof of Lemma 4

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

# Proof of Lemma 4

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

# Proof of Lemma 4

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

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

# 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

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

# Proof of Lemma 4

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

# Proof of Lemma 4

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

# Proof of Theorem 4

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

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

# 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

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

[MK18, ZG18, BM19, GK16, ZYG22]

# Proof of Theorem 4

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

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

# Proof of Theorem 4

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

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

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

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

# Proof of Theorem 4

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

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

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

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

$$\begin{aligned} \theta &\leftarrow \lambda \theta / \|\theta\|_* \\ &\Downarrow \\ \lambda &= \|\theta\|_* \end{aligned}$$

# Proof of Theorem 4

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

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

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

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

$$\lambda \geq 0$$

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

$$\boxed{\begin{array}{l} \theta \leftarrow \lambda \theta / \|\theta\|_* \\ \downarrow \\ \lambda = \|\theta\|_* \end{array}}$$

# Conclusion

Regularization = Distributional Robustness

# Take Away

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

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

# Take Away

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

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

Step 1

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

[MK18, ZG18, BM19, GK16, ZYG22]

# Take Away

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

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

Step 1

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

Step 2

# Take Away

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

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

Step 1

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

Step 2



# References

- [BK16] J. Blanchet, Y. Kang. Sample out-of-sample inference based on Wasserstein distance. OR, 2021.
- [BKM19] J. Blanchet, Y. Kang, K. Murthy. Robust Wasserstein profile inference and applications to machine learning. JAP, 2019.
- [BM19] J. Blanchet, K. Murthy. Quantifying distributional model risk via optimal transport. MOR, 2019.
- [BMN21] J. Blanchet, K. Murthy, V. A. Nguyen. Statistical Analysis of Wasserstein Distributionally Robust Estimators. Tutorials in OR, 2021.
- [BMZ18] J. Blanchet, K. Murthy, F. Zhang. Optimal transport based distributionally robust optimization: Structural properties and iterative schemes. MOR, 2021.
- [CP18] R. Chen, I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. JMLR, 2018.
- [Gao20] R. Gao. Finite-Sample Guarantees for Wasserstein Distributionally Robust Optimization: Breaking the Curse of Dimensionality. arXiv, 2020.
- [GCK17] R. Gao, X. Chen, A. J. Kleywelt. Wasserstein distributionally robust optimization and variation regularization. arXiv, 2017.
- [GK16] R. Gao, A. J. Kleywelt. Distributionally robust stochastic optimization with Wasserstein distance. arXiv, 2016.
- [GSS15] I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples. In ICLR, 2015
- [H-NW20] N. Ho-Nguyen, S. J. Wright. Adversarial Classification via Distributional Robustness with Wasserstein Ambiguity. MP, 2022.
- [MK18] P. Mohajerin Esfahani, D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric. MP, 171(1- 2):115–166, 2018.
- [S01] A. Shapiro. On duality theory of conic linear problems. In Semi-infinite Programming. Springer, 2001.
- [SADK22] S. Shafieezadeh-Abadeh, L. Aolaritei, F. Droller, D. Kuhn. Optimal Transport Based Distributionally Robust Optimization: Nash Equilibrium and Regularization, Working Paper, 2022.
- [SKM19] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via Mass Transportation. JMLR, 2019.
- [SMK15] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In NeurIPS 2015.
- [TPJR18] Y. Tian, K. Pei, S. Jana, and B. Ray. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In ICSE, 2018.
- [TSK2022] B. Taskesen, S. Shafieezadeh-Abadeh, D. Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution, MP, 2022.
- [ZG18] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. ORL, 2018.
- [ZYG22] L. Zhang, J. Yang, and R. Gao. A Simple Duality Proof for Wasserstein Distributionally Robust Optimization. arXiv, 2022.