Distributionally Robust Universal Classification

Georgia Tech College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

Authors
Siyuan Chen and Weijun Xie

Presenter
Siyuan Chen

Universal Classification

Let $Y \in \mathcal{Y} = [m]$ be the class label and $\boldsymbol{X} \in \mathcal{X} \in \mathbb{R}^d$ be the feature vector. The universal classifier space contains all Lebesgue measurable mappings $f : \mathcal{X} \to \mathcal{Y}$:

$$v_{UC} = \inf_{f \in \mathcal{F}} \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathbb{P}_0} \left[\mathbb{1}\{f(\boldsymbol{X}) \neq Y\}\right].$$

In practice, $\mathbb{P}_0$ is unknown and we only observe samples $\{(\widehat{\boldsymbol{x}}^i, \widehat{y}^i)\}_{i=1}^n$. We hope to develop a data-driven method to approximate the theoretical value v_{UC}.

Distributionally Robust UC

We define a type-1 Wasserstein ambiguity set centered at the empirical distribution $\widehat{\mathbb{P}}_n$:

$$\mathcal{P}_W(\mathcal{X} \times \mathcal{Y}) = \left\{\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \ W_1(\mathbb{P}, \widehat{\mathbb{P}}_n) \leq \theta\right\},$$

where the transportation cost is

$$c\big((\boldsymbol{x}, y), (\boldsymbol{x}', y')\big) = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa \mathbb{1}\{y \neq y'\}.$$

The (global) DRUC problem is

$$v_n^{\mathcal{X}} = \inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}_W(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbb{P}} \left[\mathbb{1}\{f(\boldsymbol{X}) \neq Y\}\right].$$

In-sample DRUC

To overcome the infinite-dimensional nature of global DRUC, we restrict the perturbed distribution to be supported on the in-sample feature space $\widehat{\mathcal{X}}_n \times \mathcal{Y}$. That is,

$$v_n^{\widehat{\mathcal{X}}_n} = \inf_{f \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{P}_W(\widehat{\mathcal{X}}_n\times\mathcal{Y})} \mathbb{E}_{\mathbb{P}} \left[\mathbb{1}\{f(\boldsymbol{X}) \neq Y\}\right].$$

Between In-sample and global DRUC.

- We consider the nearest-neighbor extension $\widetilde{f}$ to connect an in-sample classifier $\widehat{f}$ to global solutions.
- Through nearest-neighbor extension, we prove that the in-sample DRUC is asymptotically equivalent to DRUC.

Consistency Guarantee

Theorem (Informal). *There exists a positive function $\epsilon(\theta)$ depending only on the Wasserstein radius θ such that*

$$v_{UC} \leq \lim_{n\to\infty} v_n^{\widehat{\mathcal{X}}_n} = \lim_{n\to\infty} v_n^{\mathcal{X}} \leq v_{UC} + \epsilon(\theta), \qquad \lim_{\theta \downarrow 0} \epsilon(\theta) = 0.$$

Interpretation.

- Both formulations are consistent with the universal classification v_{UC} as $\theta \downarrow 0$ and sample size goes to infinity.

MILP Reformulation

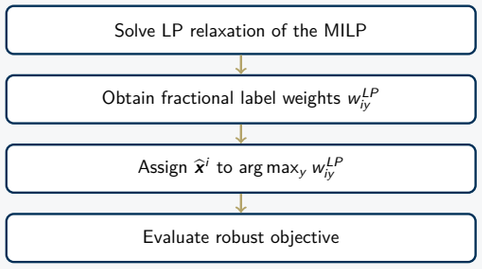
Let $d_{ij} = \|\widehat{\boldsymbol{x}}^i - \widehat{\boldsymbol{x}}^j\|$. The in-sample DRUC value $v_n^{\widehat{\mathcal{X}}_n}$ equals the optimal value of

$$\begin{aligned}
\min \quad & \theta\alpha + \frac{1}{n}\sum_{i\in[n]} \lambda_i \\
\text{s.t.} \quad & d_{ij}\alpha + \lambda_i + w_{j\widehat{y}^i} \geq 1, && \forall i, j \in [n], \\
& \kappa\alpha + \lambda_i \geq 1, && \forall i \in [n], \\
& \sum_{y\in[m]} w_{iy} = 1, && \forall i \in [n], \\
& \alpha \geq 0,\ \lambda_i \geq 0,\ w_{iy} \in \{0,1\}, && \forall i \in [n],\ y \in [m].
\end{aligned}$$

$w_{iy} = 1$ assigns label y to in-sample feature point $\widehat{\boldsymbol{x}}^i$.

MaxLin Approximation Algorithm

MILP is computationally expensive as n grows. We propose **MaxLin**:



Theorem (Approximation Ratio). *Let $\{\widetilde{w}_{iy}\}_{i\in[n],y\in[m]}$ be the solution obtained through MaxLin, and its corresponding objective value in the MILP be $\widetilde{v}_n$. Then we have $v_n^{\widehat{\mathcal{X}}_n} \leq \widetilde{v}_n \leq 2v_n^{\widehat{\mathcal{X}}_n}$.*

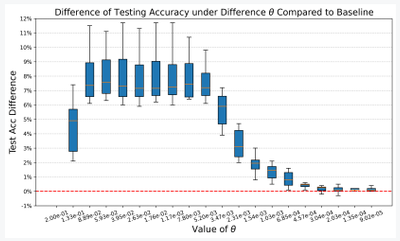
Approximation of MaxLin

We design three synthetic datasets with $n = 200$:

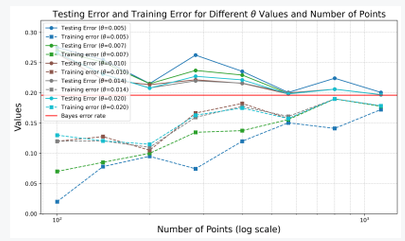
- MaxLin objective value is typically within about **5–11%** of the best MILP lower bound.
- Testing accuracies are close to MILP solutions.
- MaxLin is substantially faster, especially in harder parameter regimes.

Dataset	Obj. Gap	MaxLin Acc.	MILP Acc.	MaxLin Time (s)	MILP Time (s)
Gaussian	5.6–11.4%	73.6–77.8%	74.6–78.6%	0.10–0.29	0.36–19.9
Two-Corner	4.9–11.5%	76.0–78.8%	77.7–78.9%	0.12–0.34	0.31–18.6
Wave	5.1–9.0%	73.2–77.2%	74.7–76.4%	0.11–0.35	0.35–59.3

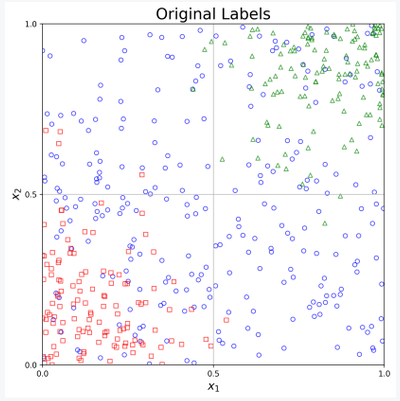
Illustration of the Classification Effect



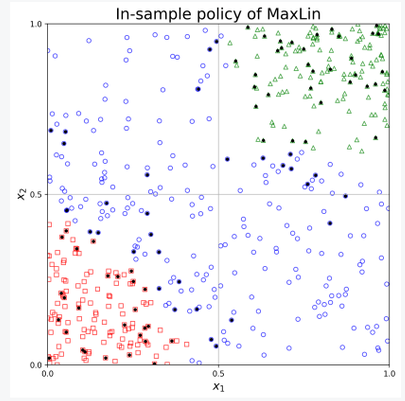
Accuracy improvement vs. θ.



Training/testing error vs. sample size.



Original labels.



Optimal in-sample classifier by MaxLin.

- MaxLin improves testing accuracy for a suitable robustness radius θ and yields a stable classifier with good performance.
- The in-sample classifier refines the original labels in uncertain regions, illustrating how the method exploits local structure in the feature space.

Correspondence: Siyuan Chen, Georgia Institute of Technology