Distributionally Robust Hypothesis Tests

Yao Xie

School of Industrial and Systems Engineering
Georgia Institute of Technology

August 2022

IFDS Workshop on Distributional Robustness in Data Science

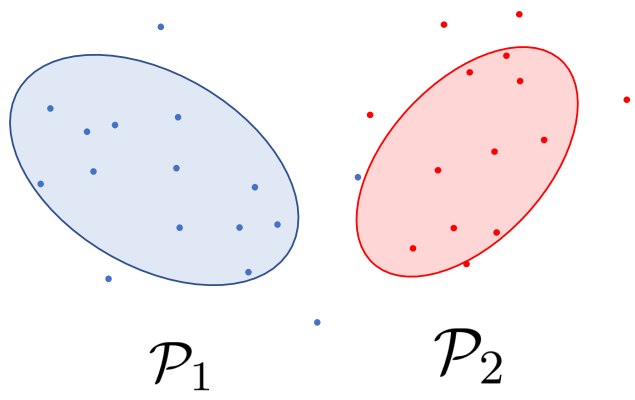
Agenda

- Background
- Distributionally robust test: Wasserstein sets
- Distributionally robust test: Sinkhorn sets
- Extensions: classification, online detection

Hypothesis test

- A general framework for *decision-making under uncertainty*
- $\Omega \subset \mathbb{R}^D$: observation space
- $\mathcal{P}_1, \mathcal{P}_2$: uncertainty sets for two families of distributions
- **Test** or **Detector**: function $\phi : \Omega \to \{1, 2\}$.
- Goal: Design a test ϕ to minimize "errors" on test data

$$
\begin{aligned}
H_1 \ &: \quad x \sim P_1, \quad P_1 \in \mathcal{P}_1, \\
H_2 \ &: \quad x \sim P_2, \quad P_2 \in \mathcal{P}_2.
\end{aligned}
$$



Simple hypothesis test

- Neyman-Pearson (1933)

$$H_1 \ : \ x \sim P_1$$
$$H_2 \ : \ x \sim P_2$$

- Optimal test is likelihood-ratio based
- Given a test sample x, decide H_2 when $\phi(x) = \frac{P_2(x)}{P_1(x)} \geq b$.
- ϕ minimizes Type-II error $\mathbb{P}_{P_2}(\phi(x) < b)$, among all test satisfies Type-I error constraint $\mathbb{P}_{P_1}(\phi(x) < b) < \alpha$.


Parameter and distributional uncertainty

We never really know the true distributions P_1 and P_2, and they have to be determined from data to get

$$\hat{P}_1, \hat{P}_2$$

Parameter uncertainty

- Parameters of P_1 and P_2 (θ_1 and θ_2) are unknown.

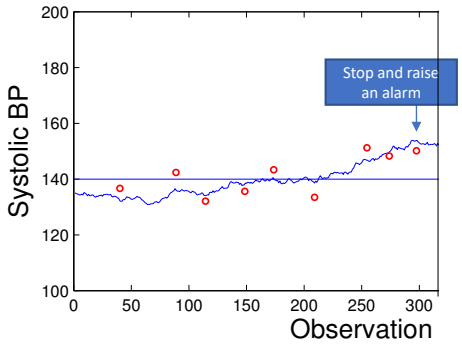
Distributional uncertainty

- Distributional form of p_1 and p_2 are not sure.

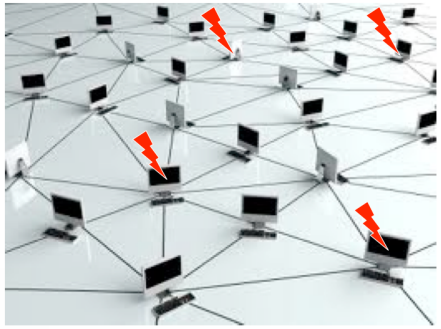
Inference with limited samples

- We may not have enough sample to obtain good estimate the distribution (or their parameters)
- How to account for **uncertainty** when design detectors?

- **Quick change-point detection:** blood pressure monitoring using wearable sensor data

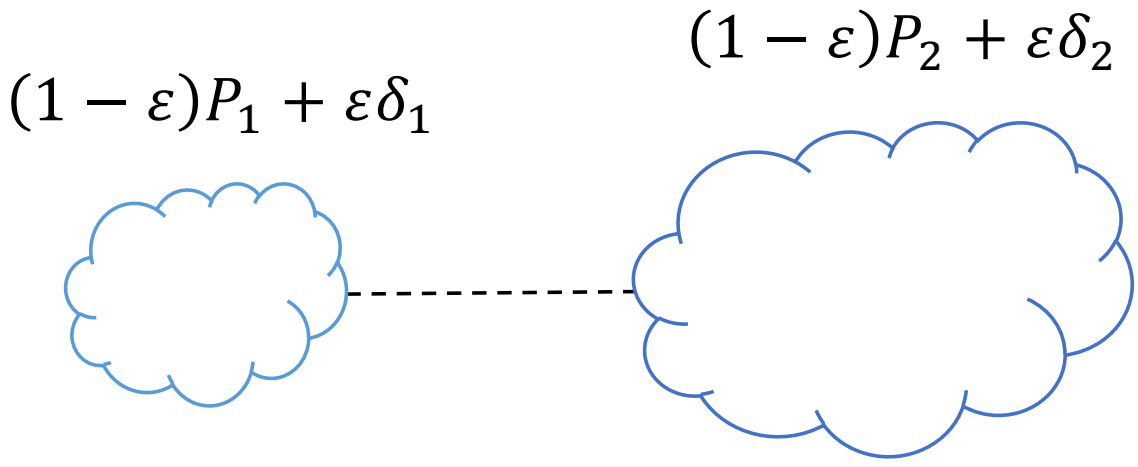


- **Anomaly detection:** network intrusion detection: lots of "normal" data, but very few anomalous data and new anomalous pattern can appear



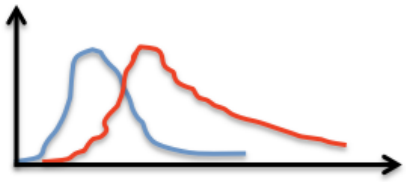
Classic robust hypothesis test

- ϵ-contamination set (Huber, 1965)
- Minimax optimal test using **least favorable distributions (LFDs)**: likelihood ratio test for LFDs from two general sets
- Conditions requiring "stochastic order" of distributions

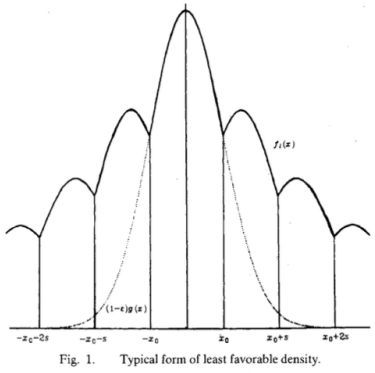


What's the difficulty?

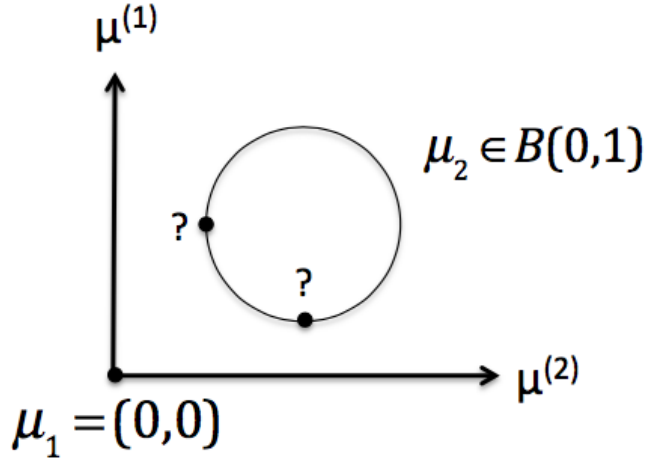
- Huber's ϵ-contamination sets hard to analyze
- In general, hard to determine LFDs in multi-dimensional cases
- Stochastic order not well-posed in multi-dimensional cases



Scalar case



least favorable density
(Moustakides, 1985)



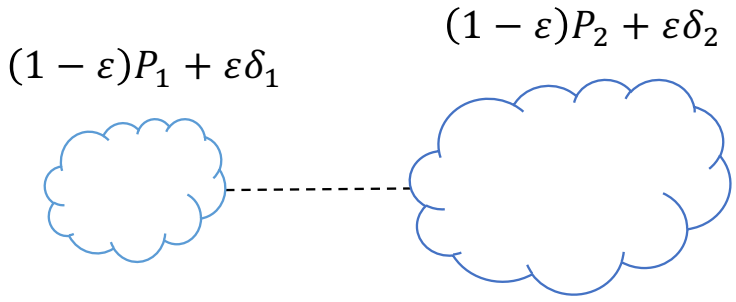
Multivariate case

X_1 stochastically larger than X_2 only when $\Sigma_1 = \Sigma_2$, and $\mu_1 \geq \mu_2$ element-wise.

Classical versus new

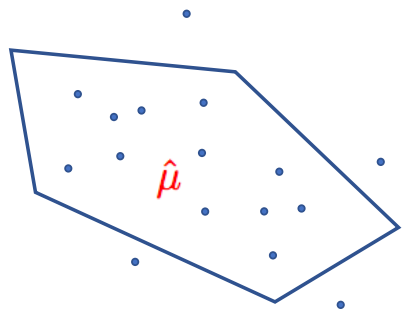
Classic robust hypothesis test (Huber 1965)

- ϵ-contamination set
- Hard to compute; intractable in high-dimensions



Robust optimization based (Goldenshluger, Juditsky, Nemirovski, 2012)

- Hypothesis test using **parametric** uncertainty sets
- cast into robust optimization



Related to robust optimization

SIAM J. Optim.
Vol. 27, No. 4, pp. 2258–2275

DISTRIBUTIONALLY ROBUST STOCHASTIC PROGRAMMING*

ALEXANDER SHAPIRO†

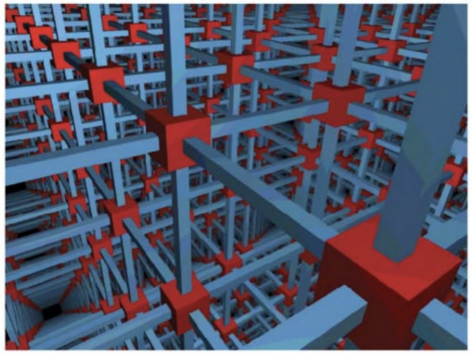
Abstract. In this paper we study distributionally robust stochastic programming in a setting where there is a specified reference probability measure and the uncertainty set of probability measures consists of measures in some sense close to the reference measure. We discuss law invariance of the associated worst case functional and consider two basic constructions of such uncertainty sets. Finally we illustrate some implications of the property of law invariance.

Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh

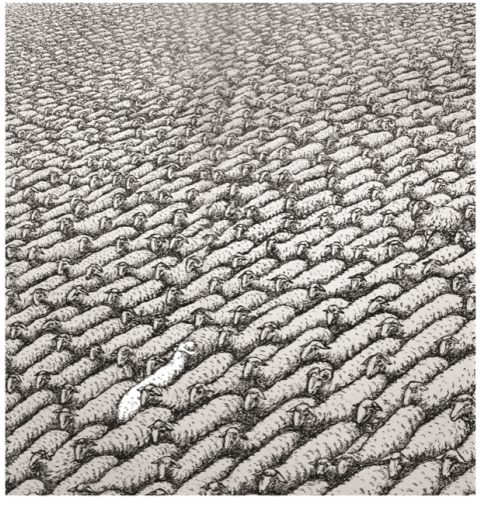
Published Online: 2 Oct 2019 | https://doi.org/10.1287/educ.2019.0198

Robust Optimization



Aharon Ben-Tal
Laurent El Ghaoui
Arkadi Nemirovski

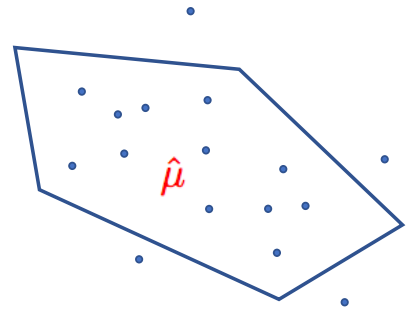
Statistical Inference via Convex Optimization



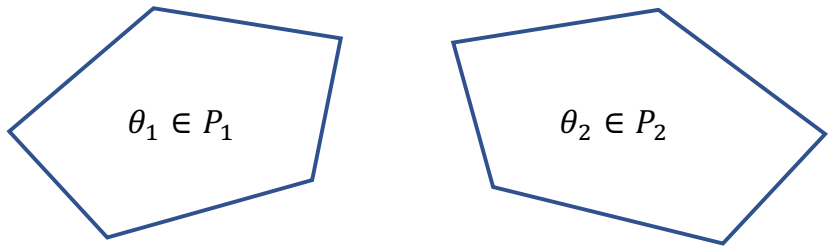
Anatoli Juditsky and
Arkadi Nemirovski

Uncertainty set for distribution parameters

- Construct uncertainty set for **parameters** of distribution



$$\begin{aligned} \mathsf{H}_1: &\quad x \sim f_{\theta_1}, \quad \theta_1 \in \mathcal{P}_1 \\ \mathsf{H}_2: &\quad x \sim f_{\theta_2}, \quad \theta_2 \in \mathcal{P}_2 \end{aligned}$$

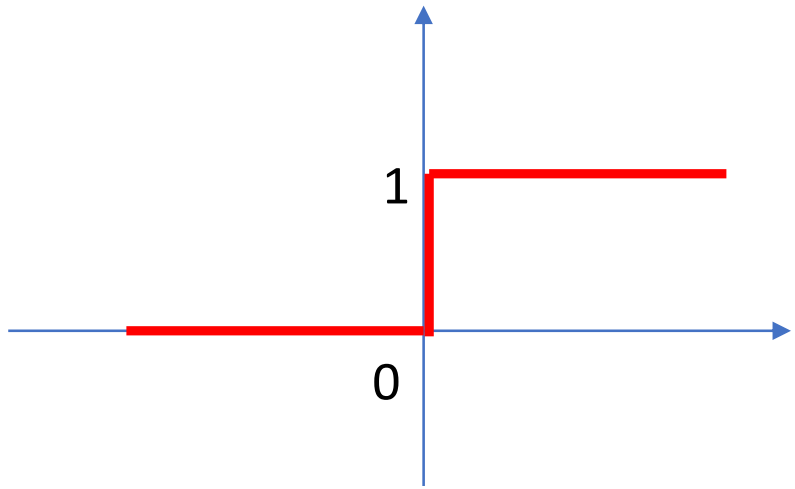


Hypothesis test using convex optimization

(Goldenshluger, Juditsky and Nemirovski 2015)

- Single observation $x \in \mathbb{R}^D$
- Detector $\phi : \mathbb{R}^D \to \mathbb{R}$: Claim H_1 if $\phi(x) > 0$
- Optimal ϕ: minimize worst-case type-I and type-II error

$$\begin{aligned} & \frac{1}{2}\mathbb{P}_{f_1}(\phi(X) > 0) + \frac{1}{2}\mathbb{P}_{f_2}(\phi(X) < 0) \\ = \quad & \frac{1}{2}\mathbb{E}_{f_1}[\mathbb{I}\{\phi(X) > 0\}] + \frac{1}{2}\mathbb{E}_{f_2}[\mathbb{I}\{\phi(X) < 0\}] \end{aligned}$$

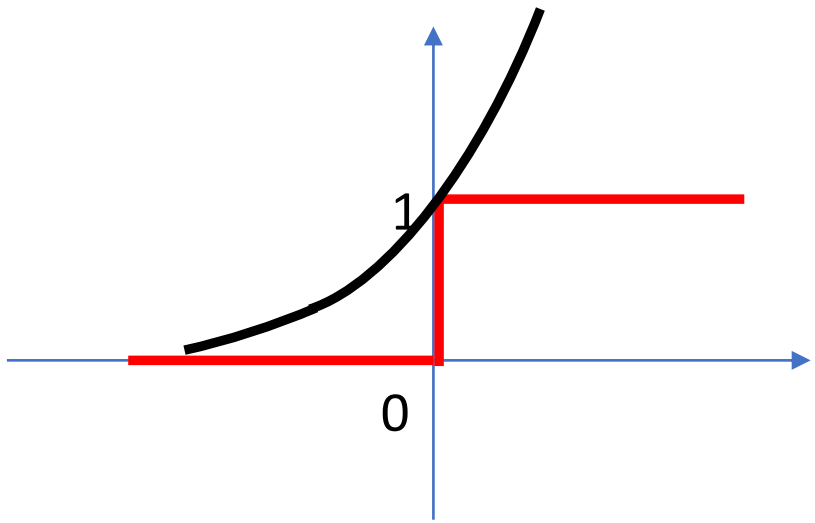


- Use **exponential loss** as an relaxation to 0-1 loss

$$\Phi(\phi, \theta_1, \theta_2) = \frac{1}{2}\int e^{-\phi(\omega)} f_{\theta_1}(\omega) d\omega + \frac{1}{2}\int e^{\phi(\omega)} f_{\theta_2}(\omega) d\omega$$

- Solve an detector that achieves minimax optimality

$$\min_{\phi} \max_{\theta_0 \in \mathcal{P}_1, \theta_1 \in \mathcal{P}_2} \Phi(\phi, \theta_1, \theta_2)$$



Result from

(Goldenshluger, Juditsky and Nemirovski 2015)

- Optimal detecter given by

$$\phi^*(x) = \log \frac{f_{\theta_2^*}(x)}{f_{\theta_1^*}(x)} \quad \text{(Neyman-Pearson like)}$$

- "Least favorable" parameters given by

$$(\theta_1^*, \theta_2^*) = \arg \max_{\theta_1 \in \mathcal{P}_1, \theta_2 \in \mathcal{P}_2} \underbrace{\int \sqrt{f_{\theta_1}(\omega) f_{\theta_2}(\omega)} d\omega}_{1-\text{Hellinger distance}}$$

- Optimal risk $\Phi(\phi^*, \theta_1^*, \theta_2^*)$

$$\text{Type-I error} < \epsilon^*, \quad \text{Type-II error} < \epsilon^*$$

Example: Gaussian mean uncertainty

- For a Gaussian case, the optimization problem is explicit

$$\begin{aligned} \mathsf{H}_1: &\quad X \sim \mathcal{N}(\mu_1, \Sigma), \quad \mu_1 \in \mathcal{P}_1 \\ \mathsf{H}_2: &\quad X \sim \mathcal{N}(\mu_2, \Sigma), \quad \mu_2 \in \mathcal{P}_2 \end{aligned}$$

- Optimal detector $\phi^* = a^\intercal \omega - b$

$$a = \frac{1}{2}\Sigma^{-1}(\mu_1^* - \mu_2^*), \quad b = \frac{1}{2}a^\intercal \Sigma^{-1}(\mu_1^* + \mu_2^*)$$

$$(\mu_1^*, \mu_2^*) = \arg \min_{\mu_1 \in \mathcal{P}_1, \mu_2 \in \mathcal{P}_2} (\mu_1 - \mu_2)^\intercal \Sigma^{-1} (\mu_1 - \mu_2)$$

- Optimal risk

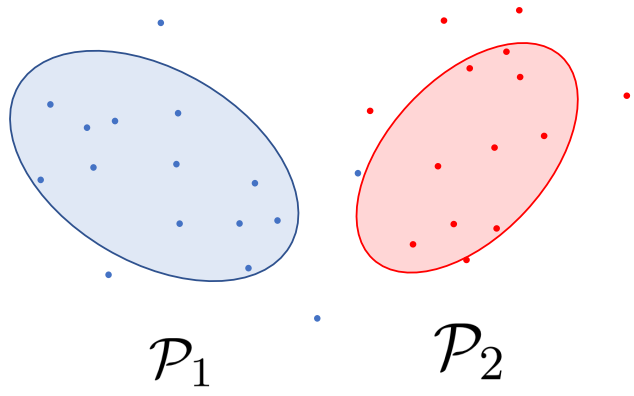
$$\epsilon^* = e^{-\frac{1}{8}(\mu_1^* - \mu_2^*)\Sigma^{-1}(\mu_1^* - \mu_2^*)}$$

Agenda

- Background
- Distributionally robust test: Wasserstein sets $\leftarrow$
  (Joint work with Liyan Xie, Rui Gao)
- Distributionally robust test: Sinkhorn sets
- Extensions: classification, online detection

Distributionally robust test

$$
\begin{aligned}
H_1 &: \quad x \sim P_1, \quad P_1 \in \mathcal{P}_1, \\
H_2 &: \quad x \sim P_2, \quad P_2 \in \mathcal{P}_2.
\end{aligned}
$$

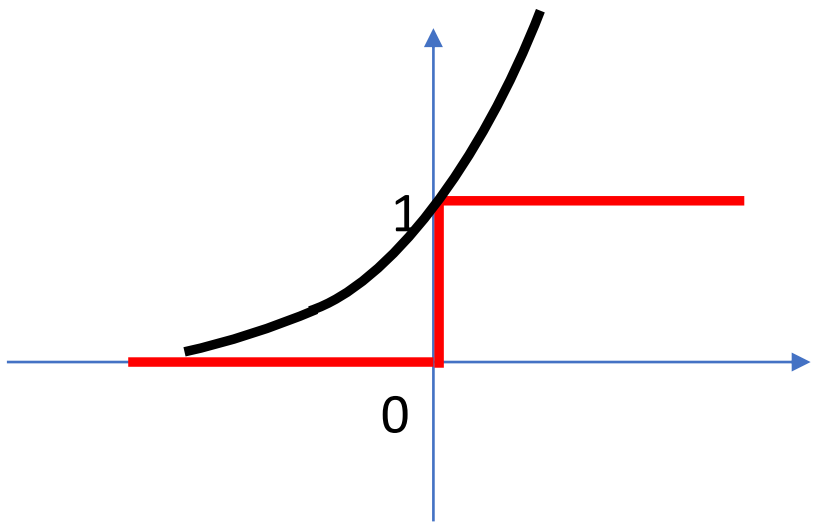


- Uncertainty sets $\mathcal{P}_1$ and $\mathcal{P}_2$ are constructed based on **empirical distributions** and induced by "distance" or "divergence"
- Distributionally robust stochastic programming (Shapiro, 2017)

Distributionally robust test: Convex relaxation

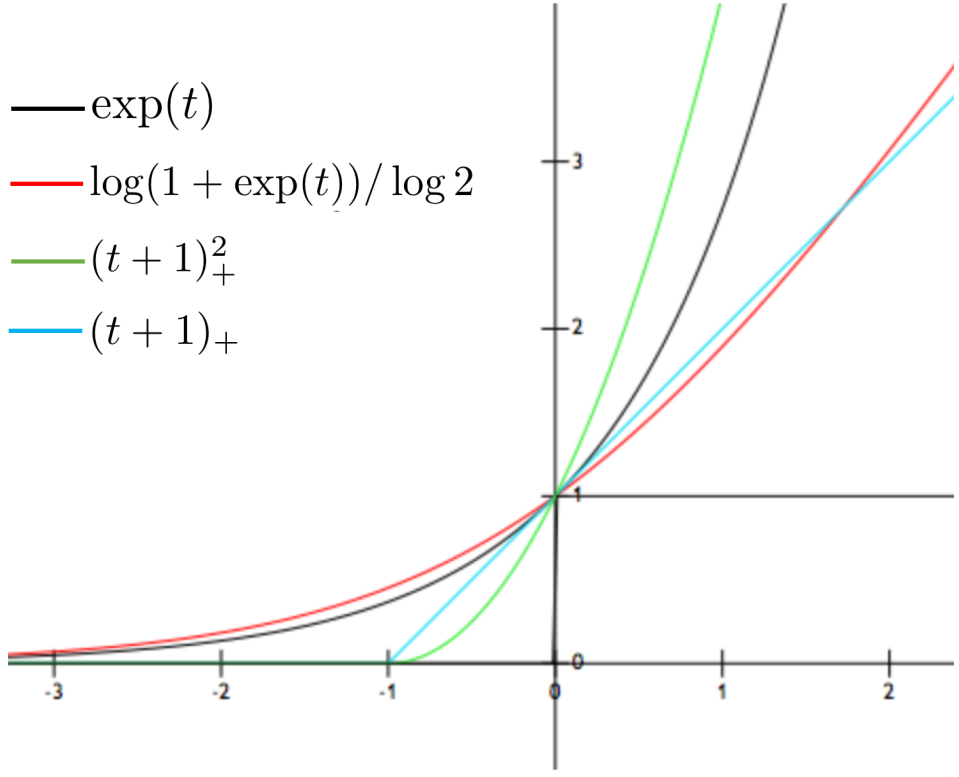
$$\inf_{\phi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\phi; P_1, P_2).$$

- Decide H_1 when $\phi(x) \geq 0$
- (Risk) $\Phi(\phi; P_1, P_2) := \mathbb{E}_{P_1}[\ell \circ (-\phi)(x))] + \mathbb{E}_{P_2}[\ell \circ \phi(\omega)]$.
- Type I error + Type II error $\leq \Phi(\phi; P_1, P_2)$



Loss function: Convex relaxation

- Loss function ℓ: non-decreasing, convex function
- $\ell(0) = 1$ and $\lim_{t \to -\infty} \ell(t) = 0$.



Choices of uncertainty sets

For robust hypothesis test

- KL (Levy 09) (Gul, Zoubir 2017): **discrete** distributions are supported only on training samples.
- Wasserstein (Xie, Gao, X. 2019): contain both **discrete** and **continuous** distributions, least-favorable distributions (LFD) are supported only on training samples.
- Sinkhorn: LFD can be **continuous** distributions due to entropic regularization.

Wasserstein uncertainty sets

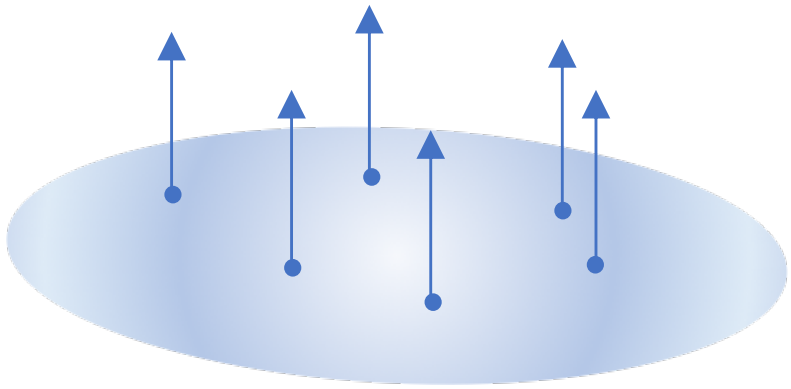
- Consider two sets of samples $\{x_i^k\}$, $i = 1, \ldots, n_k$, $k = 1, 2$
- Empirical distributions

$$\hat{P}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \delta_{x_i^k}, \quad k = 1, 2$$

- Construct uncertainty sets

$$\mathcal{P}_k = \{P : \mathcal{W}(P, \hat{P}_k) \leq \theta_k\}, \quad k = 1, 2$$

$\theta_k > 0$: size of uncertainty set

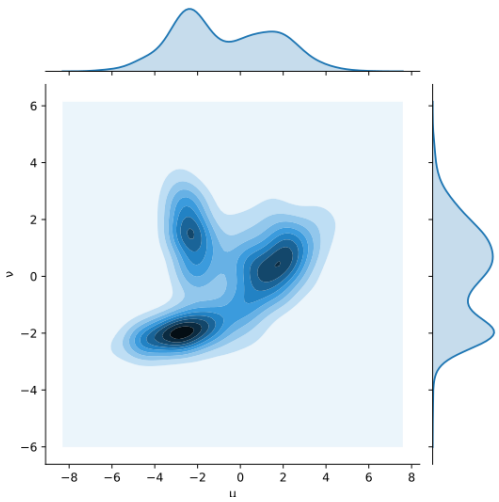


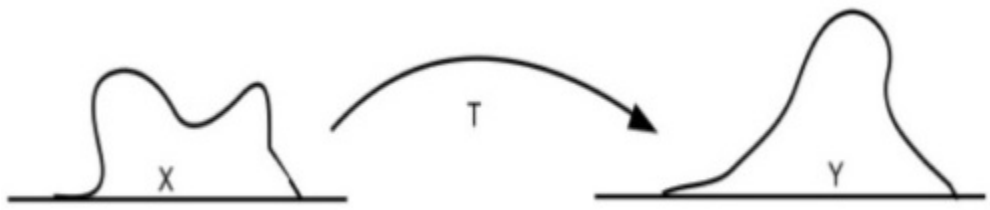
Wasserstein divergence

- Definition of Wasserstein divergence

$$\mathcal{W}(P,Q) := \min_{\gamma \in \mathscr{P}(\Omega^2)} \Big\{ \mathbb{E}_{(\omega,\omega')\sim\gamma} \left[c(\omega,\omega') \right] : \\ \gamma \text{ has mariginal distributions } P \text{ and } Q \Big\}$$

- Wasserstein divergence: also know as **earth mover distance**: minimum cost of transporting probability masses



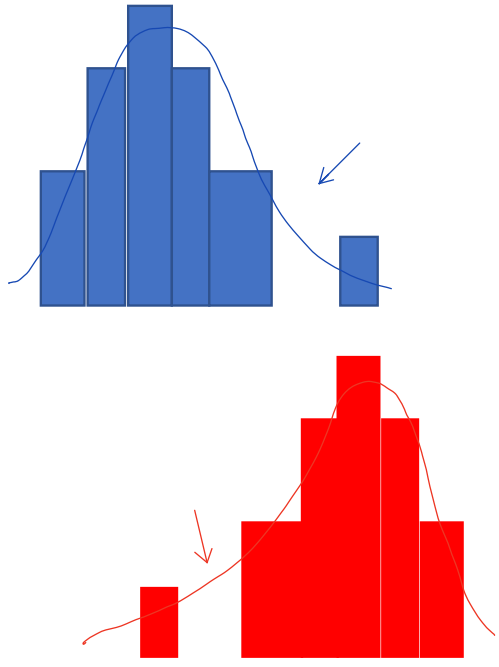


KL vs. Wasserstein

Kullback-Leibler (KL) divergence between two discrete distributions

$$\mathrm{KL}(P||Q) = \sum_{i \in \Omega} P_i \log\left(\frac{P_i}{Q_i}\right)$$

- P and Q need same support
- **Wasserstein** metric can be defined for distributions with different support; transport cost can incorporate underlying data geometry
- With **limited data**, empirical distributions may not have "gaps"
- Wasserstein uncertainty set may contain "more" distributions



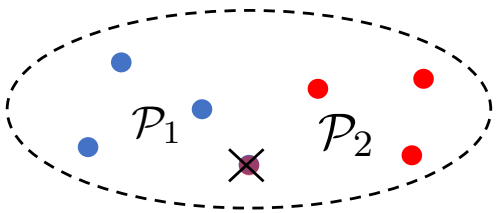
Strong duality

Consider $c(x, y) = \|x - y\|_2$

Theorem (Xie, Gao, X., 2019)

$$\inf_{\phi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\phi; P_1, P_2) = \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \inf_{\phi} \Phi(\phi; P_1, P_2).$$

Proof idea: (Structural property of least-favorable distribution)
LFDs P_1^* and P_2^* have the same supports on *the union of two empirical distributions*



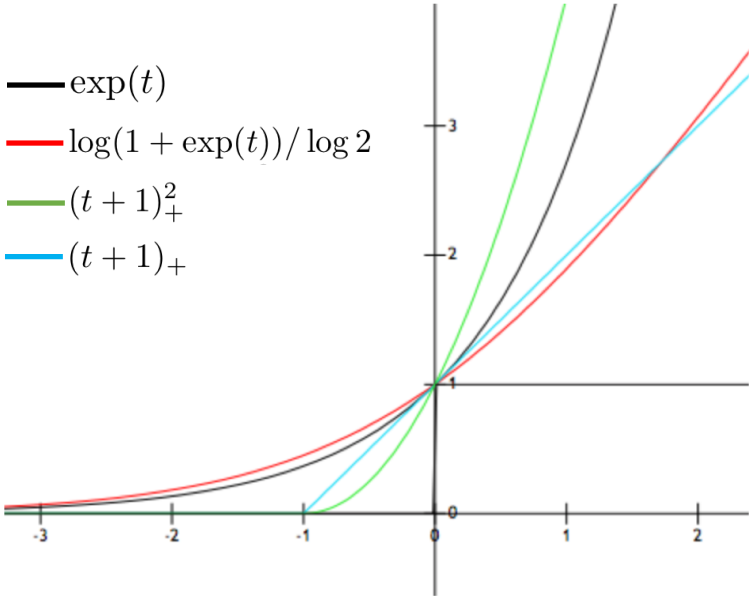
Solution strategy

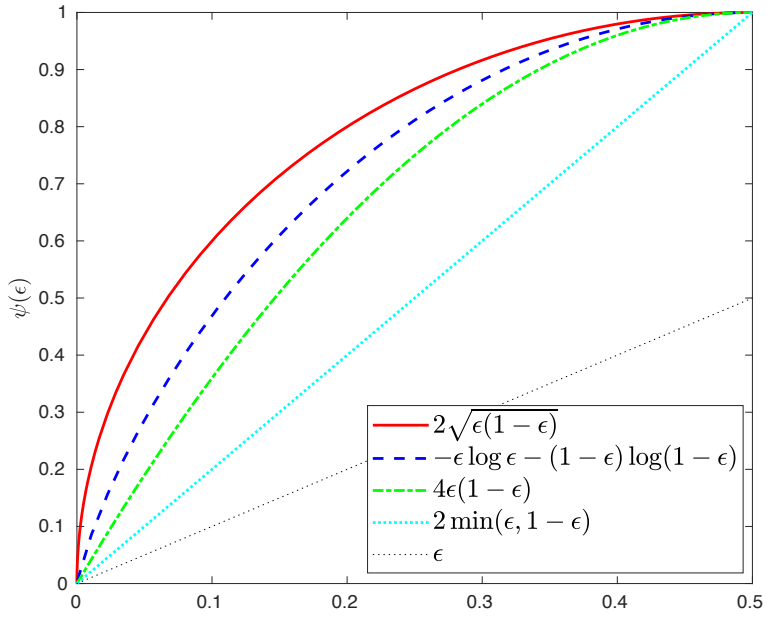
$$\inf_{\phi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\phi; P_1, P_2).$$

- Step 1: Solve **optimal detector** for a pair of distributions
- Step 2: Solve **least favorable distributions** (LFD)

$$\sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \inf_{\phi} \Phi(\phi; P_1, P_2)$$

$$= \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \int_{\Omega} \psi\left(\tfrac{dP_1}{d(P_1+P_2)}\right) d(P_1 + P_2).$$





Generating function	Auxiliary function	Optimal detector	Detector risk
$\ell(t)$	$\psi(p)$	ϕ^*	$1-1/2\inf_\phi \Phi(\phi;P_1,P_2)$
$\exp(t)$	$2\sqrt{p(1-p)}$	$\ln\sqrt{p_1/p_2}$	$H^2(P_1,P_2)$
$\log(1+\exp(t))/\log 2$	$-H(p)/\log 2$	$\log(p_1/p_2)$	$JS(P_1,P_2)/\log 2$
$(t+1)_+^2$	$4p(1-p)$	$1-2\frac{p_1}{p_1+p_2}$	$\chi^2(P_1,P_2)$
$(t+1)_+$	$2\min(p,1-p)$	$\mathrm{sgn}(p_1-p_2)$	$TV(P_1,P_2)$

Step 1: Optimal detector given P_1 and P_2

- Radon-Nikodym derivative $\frac{dP_k}{d(P_1+P_2)}$, $k = 1, 2$
- Given (P_1, P_2)

$$\inf_{\phi} \Phi(\phi; P_1, P_2) = \int_{\Omega} \psi\left(\tfrac{dP_1}{d(P_1+P_2)}\right) d(P_1 + P_2)$$

- Optimal detector

$$-\phi^*(\omega) \in \arg\min_{t\in\mathbb{R}} \left[\tfrac{dP_1}{d(P_1+P_2)}(\omega)\ell(-t) + \tfrac{dP_2}{d(P_1+P_2)}(\omega)\ell(t)\right]$$

Step 2: Least favorable distributions

- Robust detector with Wasserstein uncertainty sets reduces to a finite-dimensional convex program

$$
\begin{aligned}
\max_{\substack{p_1,p_2\\ \gamma_1,\gamma_2}} \quad & \sum_{l=1}^{n_1+n_2} \psi\big(\tfrac{p_1^l}{p_1^l+p_2^l}\big)(p_1^l+p_2^l) \\
\text{subject to} \quad & \sum_{l=1}^{n_1+n_2}\sum_{m=1}^{n_1+n_2} \gamma_k^{lm}\left\|\omega^l-\omega^m\right\| \le \theta_k,\ k=1,2, \\
& \sum_m \gamma_1^{lm} = \frac{1}{n_1},\ 1\le l\le n_1, \quad \sum_m \gamma_1^{lm} = 0,\ n_1+1\le l\le n_1+n_2, \\
& \sum_m \gamma_2^{lm} = 0,\ 1\le l\le n_1, \quad \sum_m \gamma_2^{lm} = \frac{1}{n_2},\ n_1+1\le l\le n_1+n_2, \\
& \sum_l \gamma_k^{lm} = p_k^m,\ 1\le m\le n_1+n_2,\ k=1,2.
\end{aligned}
$$

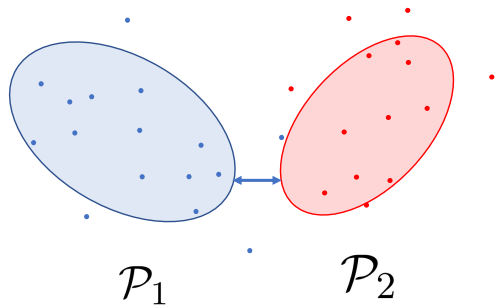
- Complexity: independent of data dimension D

Statistical interpretation

Generating function	Auxiliary function	Optimal detector	Detector risk
$\ell(t)$	$\psi(p)$	ϕ^*	$1 - 1/2 \inf_\phi \Phi(\phi; P_1, P_2)$
$\exp(t)$	$2\sqrt{p(1-p)}$	$\ln \sqrt{p_1/p_2}$	$H^2(P_1, P_2)$
$\log(1+\exp(t))/\log 2$	$-H(p)/\log 2$	$\log(p_1/p_2)$	$JS(P_1, P_2)/\log 2$
$(t+1)_+^2$	$4p(1-p)$	$1 - 2\frac{p_1}{p_1+p_2}$	$\chi^2(P_1, P_2)$
$(t+1)_+$	$2\min(p, 1-p)$	$\mathrm{sgn}(p_1 - p_2)$	$TV(P_1, P_2)$

(Juditsky, Nemirovski, 2015)

- Hellinger distance: $H^2(P_1, P_2) = 1 - \int \sqrt{p_1(x)p_2(x)}dx$
- Jensen-Shannon (JS): $JS(P_1, P_2) = \frac{1}{2}D(P_1||P_2) + \frac{1}{2}D(P_2||P_1)$
- χ^2-divergence: $\chi^2(P_1, P_2) = \int (p_1(x)/p_2(x) - 1)^2 p_2(x)dx$
- Total variation: $TV(P_1, P_2) = \sup_{A \in \mathcal{F}} |P_1(A) - Q_1(A)|$



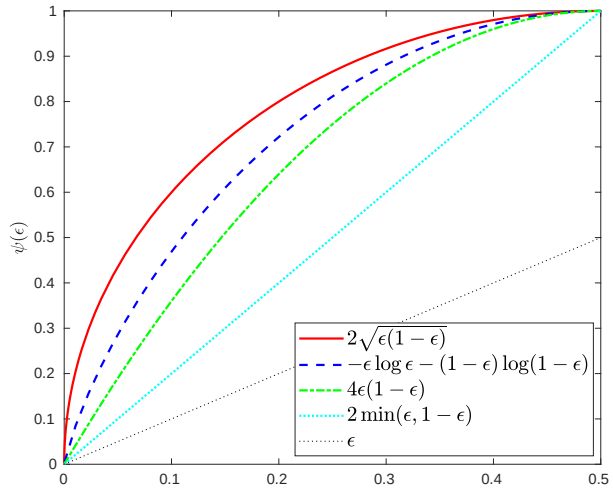
Relaxation gap

- Can prove "relaxation loss" is small
- Increase in risk due to relaxation less than $\psi(\epsilon) - \epsilon$

$$\psi(p) := \min_{t \in \mathbb{R}} \left[p\ell(t) + (1-p)\ell(-t) \right], \quad 0 \le p \le 1.$$

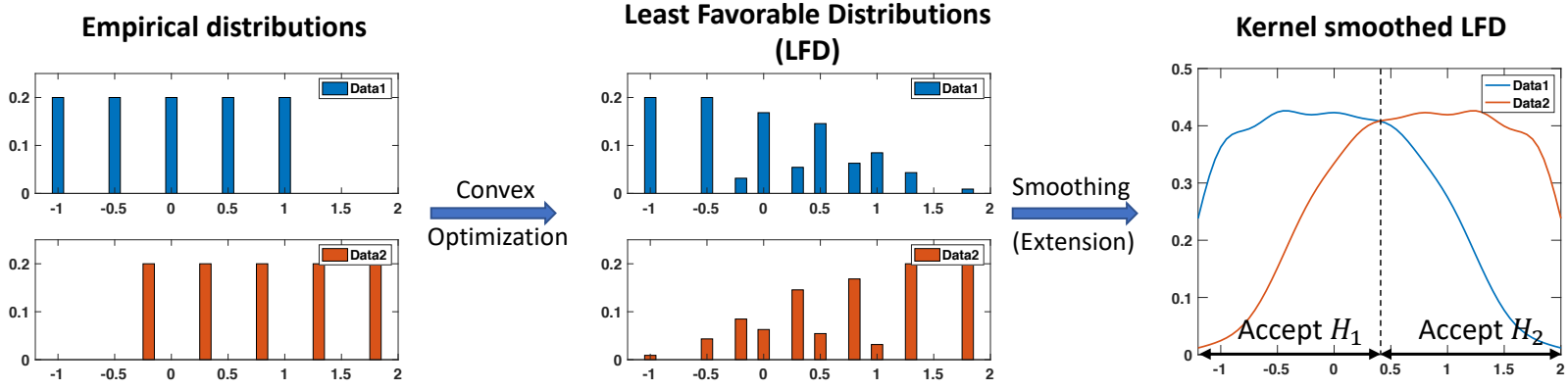
- Hinge loss has the smallest relaxation gap

$\ell(t)$	$\psi(p)$
$\exp(t)$	$2\sqrt{p(1-p)}$
$\log(1+\exp(t))/\log 2$	$-H(p)/\log 2$
$(t+1)_+^2$	$4p(1-p)$
$(t+1)_+$	$2\min(p, 1-p)$



Example: LFD

- Wasserstein uncertainty sets
- Left: Empirical distributions of two sets of training samples
- Middle: Solved LFDs
- Right: kernel smoothed versions (with kernel bandwidth 0.25)



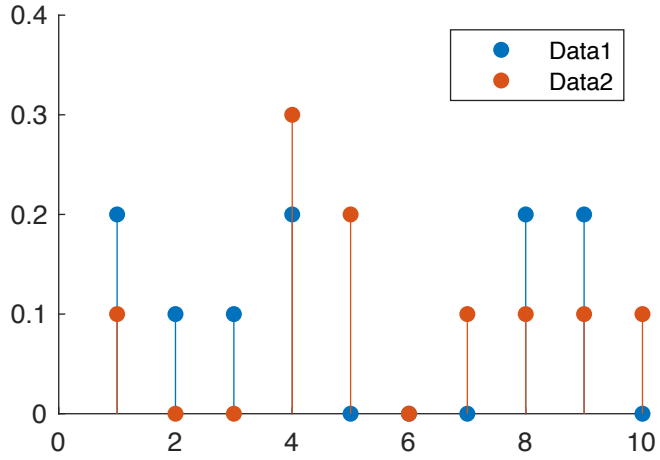
Kernel-smoothed detectors

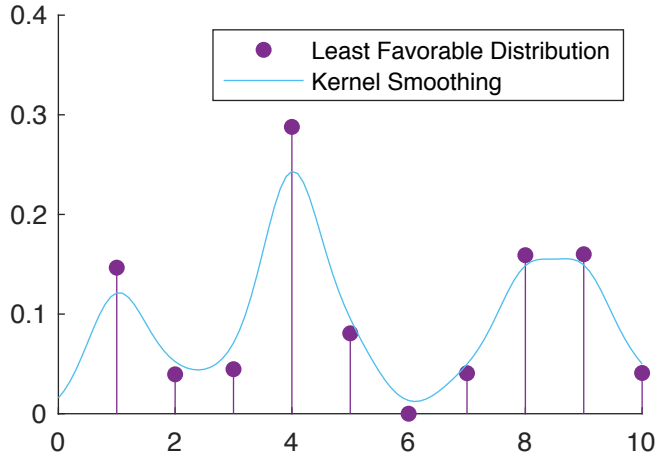
- Smoothed LFDs P_1^h and P_2^h by convolving with Gaussian kernel

$$K(x) = (2\pi)^{-D/2} e^{-\|x\|^2/(2h^2)}$$

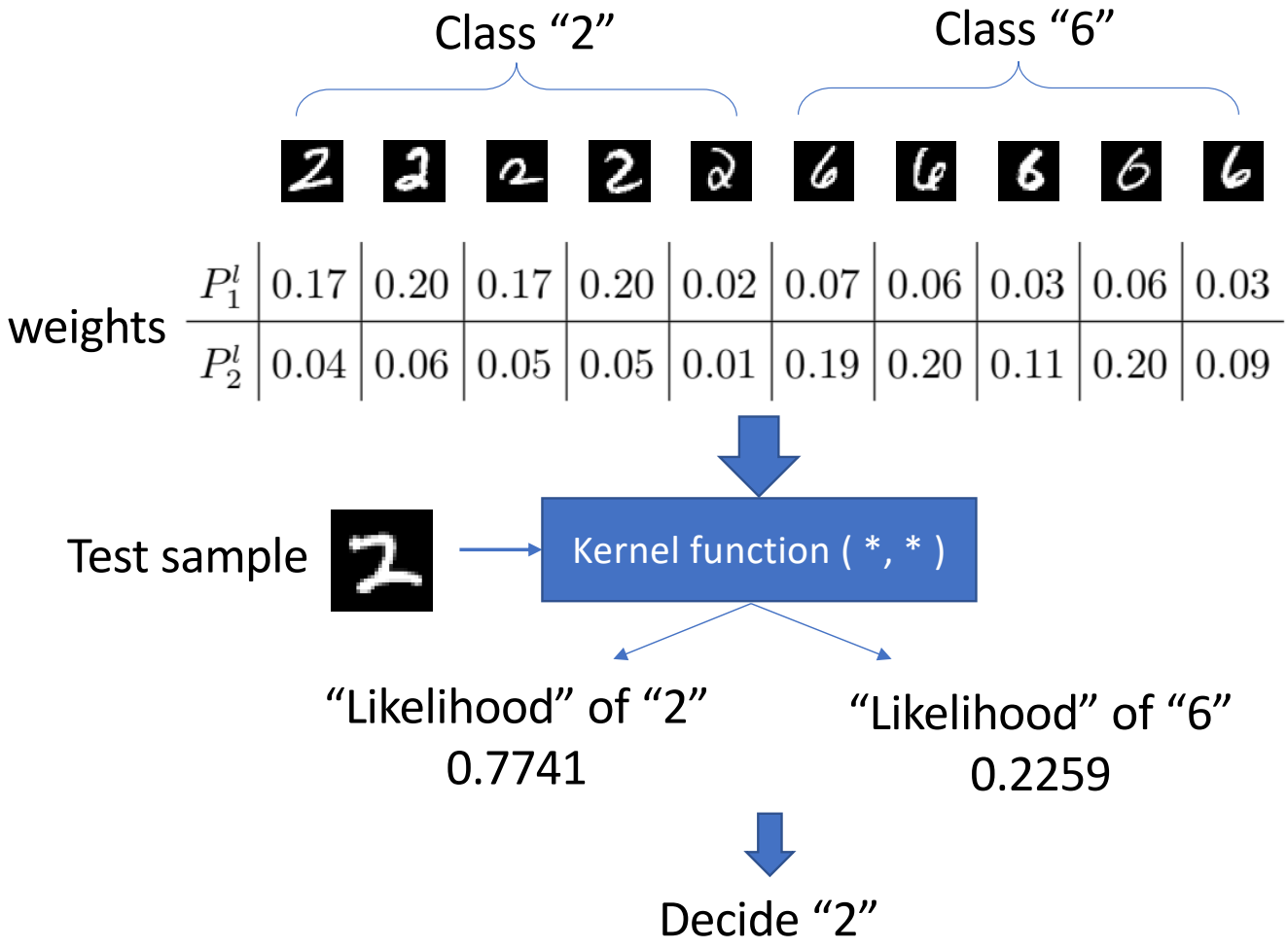
$$\left| TV(P_1^h, P_2^h) - TV(P_1^*, P_2^*) \right| \leq D(n_1 + n_2) e^{-\frac{\varrho_{\min}^2}{2Dh^2}},$$

where $\varrho_{\min} = \min_{l \neq m} \|\omega^l - \omega^m\|/2$



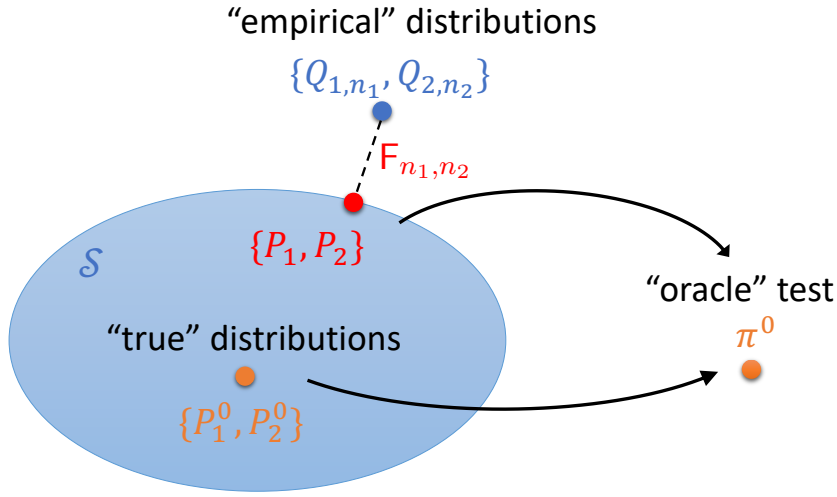


Results: MNIST dataset



How to select radius of uncertainty sets?

- Suppose global solutions (P_1^*, P_2^*, ϕ^*)
- Good radius r induces uncertainty sets that contain a pair of distributions whose optimal test coincides with the optimal test for true distributions.



- When the sample sizes for two hypotheses are balanced $\mathcal{O}(n)$

$$r \sim O(n^{-1/D})$$

"Batch" test samples

- Given n' test samples $\omega_1, \omega_2, \ldots, \omega_{n'}$
- ϕ^h detector with kernel smoothed LFDs
- Batch detector

$$\tilde{\phi}(\omega_1, \omega_2, \ldots, \omega_{n'}) = \frac{1}{n'} \sum_{i=1}^{n'} \phi_h^*(\omega_i),$$

- For ℓ being Hinge loss function

$$\text{type-I} + \text{type-II errors} \ \leq \exp\left\{-n' \cdot \frac{(1 - \Delta(P_1^h, P_2^h))^2}{2}\right\}$$

where
$\Delta(P_1^h, P_2^h) = \max\left\{\mathbb{E}_{P_1^h}[\ell \circ (-\phi^h)(\omega)], \mathbb{E}_{P_2^h}[\ell \circ (\phi^h)(\omega)]\right\}.$

Simulation example

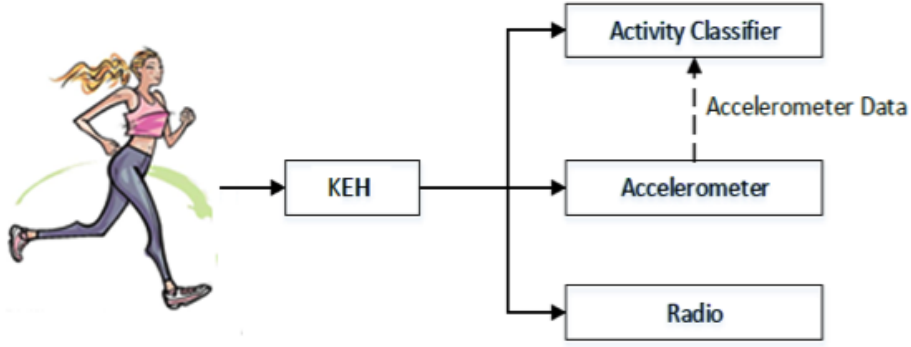
- dimension 100, test two hypotheses: Gaussian mixture models $0.5\mathcal{N}(0.4\mathbf{1}, I_{100}) + 0.5\mathcal{N}(-0.4\mathbf{1}, I_{100})$ and $0.5\mathcal{N}(0.4\mathbf{f}, I_{100}) + 0.5\mathcal{N}(-0.4\mathbf{f}, I_{100})$, $\mathbf{f} \in \mathbb{R}^{100} = [1, \ldots, 1, -1, \ldots, -1]^T$
- few training samples $n_1 = n_2 = 10$, test sample size 1000
- radius of uncertainty sets and kernel bandwidth by cross-validation
- Performance gain for small samples

TABLE 1
GMM data, 100-dimensional, comparisons averaged over 500 trials.

# observation (m)	Ours	GMM	Logistic	Kernel SVM	3-layer NN
1	**0.2145**	0.2588	0.4925	0.3564	0.4164
2	**0.2157**	0.2597	0.4927	0.3581	0.4164
3	**0.1331**	0.1755	0.4905	0.3122	0.3796
4	**0.1329**	0.1762	0.4905	0.3129	0.3808
5	**0.0937**	0.1310	0.4888	0.2877	0.3575
6	**0.0938**	0.1315	0.4881	0.2893	0.3570
7	**0.0715**	0.1034	0.4880	0.2727	0.3399
8	**0.0715**	0.1038	0.4876	0.2745	0.3401
9	**0.0579**	0.0850	0.4873	0.2634	0.3264
10	**0.0578**	0.0851	0.4874	0.2641	0.3267

Real data example: Human activity detection

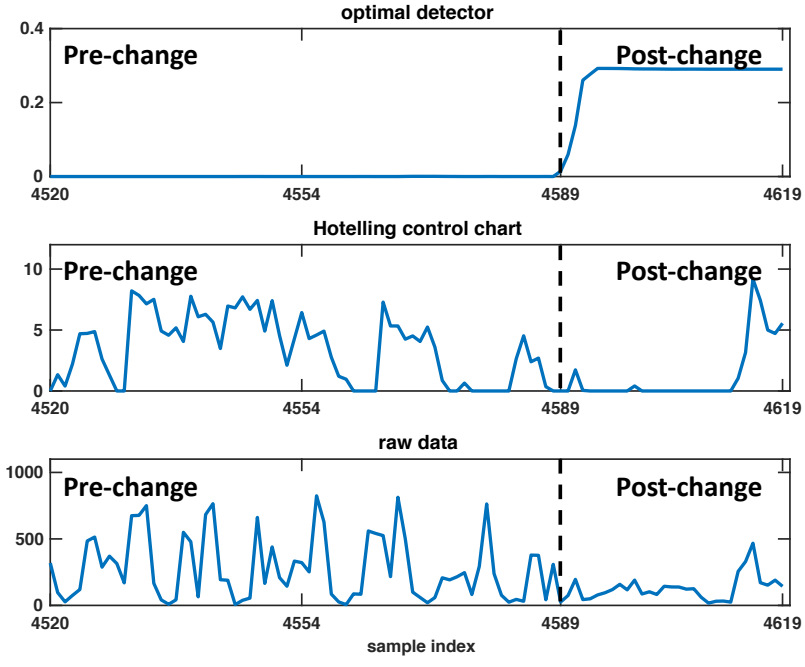
Human activity detection

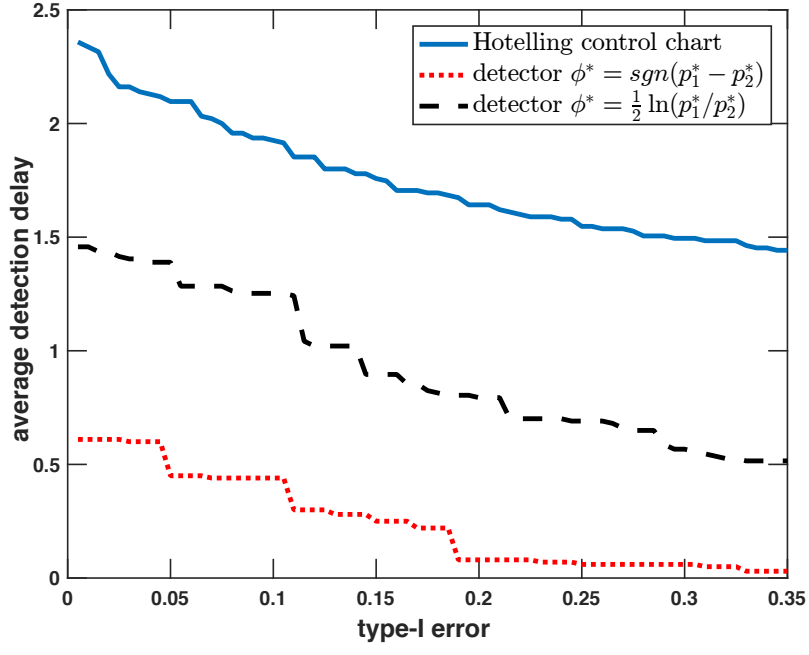


Credit: CSIRO Research

- Human Activity Sensing Consortium (HASC) challenge 2011
- Data consists of human activity information collected by portable three-axis accelerometers (6-dimensional sequential).
- Record 6 kinds of human activities:
  walk/jog, stairUp/stairDown, elevatorUp/elevatorDown, escalatorUp/escalatorDown, movingWalkway, stay.

Comparison with classical statistical method





pre-change activity jogging and post-change activity walking

Agenda

- Background
- Distributionally robust test: Wasserstein sets
- Distributionally robust test: Sinkhorn sets $\leftarrow$
  (Joint works with Jie Wang, Rui Gao, and Xiuyuan Cheng)
- Extensions: classification, online detection

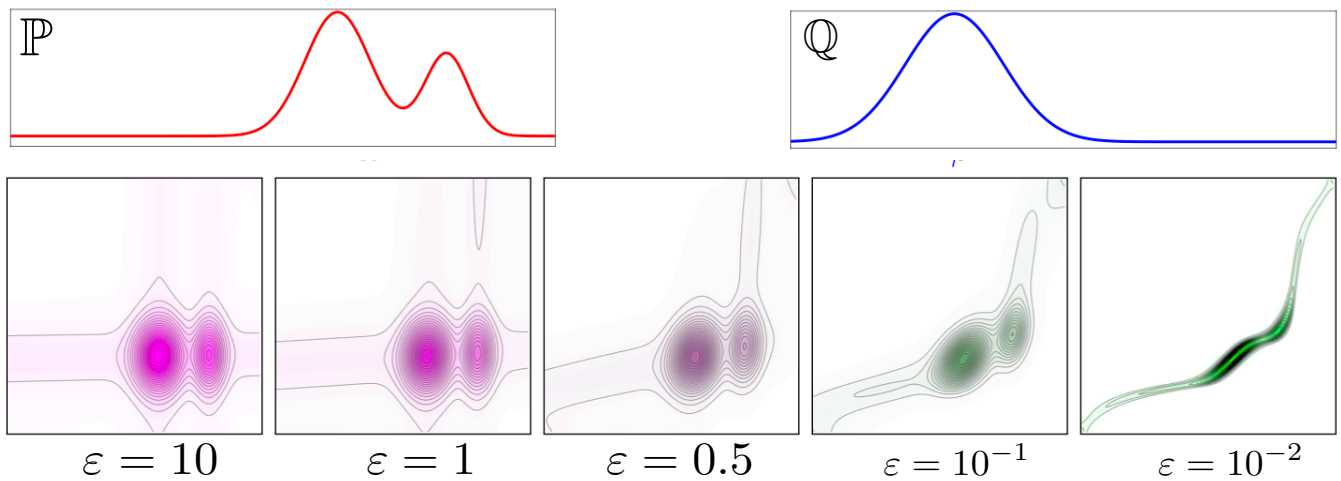
Sinkhorn Problem

- Sinkhorn problem: Entropy regularized optimal transport (Wilson 1969) (Cuturi 2013):

$$S_\epsilon(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \left\{ \mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] + \epsilon H(\gamma \mid P \otimes Q) \right\}.$$

- Relative entropy between γ and $P \otimes Q$:

$$H(\gamma \mid P \otimes Q) = \int \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}P(x)\,\mathrm{d}Q(y)}\right) \mathrm{d}\gamma(x,y).$$



Interpretation of entropic regularization

- Minimize transport cost + mutual information between two random variables through coupling

$$S_\epsilon(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \left\{ \mathbb{E}_{(X,Y)\sim\gamma}[c(X,Y)] + \epsilon I(X;Y) \right\}.$$

- $\epsilon \to \infty$, $I(X;Y) = 0$
  $(X,Y) \sim P \otimes Q =$ they are independent

- $\epsilon \to 0$ converge to Wasserstein

- Efficient computation by Sinkhorn-Knopp (Altschuler, Niles-Weed, and Rigollet 2017)

Sinkhorn Distributionally-Robust Optimization (DRO)

- Minimize worst-case loss function against the LFD

$$\min_{\theta} \max_{S_\epsilon(\hat{P},P)\leq r} \mathbb{E}_{z\sim P}[\ell_\theta(z)]$$

  $\hat{P}$ empirical distribution of n training samples $\{x_1, \ldots, x_n\}$

- Strong duality holds for convex loss function ℓ_θ: interchange min and max

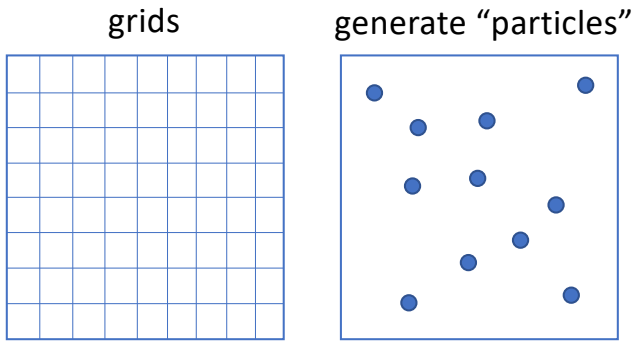
- Motivations: Robust regression, robust logistic regression, robust hypothesis testing

Discretize Sinkhorn DRO

Consider LFD problem (inner maximization):

$$\max_{S_\epsilon(\hat{P},P)\leq r} \mathbb{E}_{z\sim P}[\ell_\theta(z)]$$

- If discretize P (distribution of z) with m atoms



- Problem structure leads to highly efficient computable solution

"Lightspeed" computation

- Modified matrix scaling — optimal regularized transport:

$$\gamma_{ij}^* = \alpha_i e^{-\frac{c(x_i, z_j)}{\epsilon}} e^{\frac{\ell_\theta(z_j)}{\lambda^* \epsilon}}$$

  where α_i is a normalizing constant such that $\sum_{j=1}^m \gamma_{ij} = 1/n$
- $\{\hat{P}_i\}$, $i = 1, \ldots, n$, empirical distribution
- LFD $P_j^* = \sum_{i=1}^n \gamma_{ij}^*$, $j = 1, \ldots, m$
- Reduce to simple dual variable search

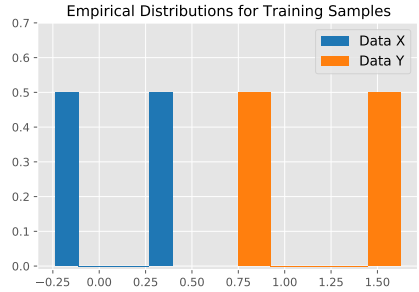
$$\lambda^* = \arg\min_{\lambda > 0} \{\lambda r + \lambda \epsilon \sum_{i=1}^n \hat{P}_i \log \left(\sum_{j=1}^m \frac{1}{\hat{P}_i} e^{\frac{\ell_j - \lambda c_{ij}}{\lambda \epsilon}} \right) \}$$

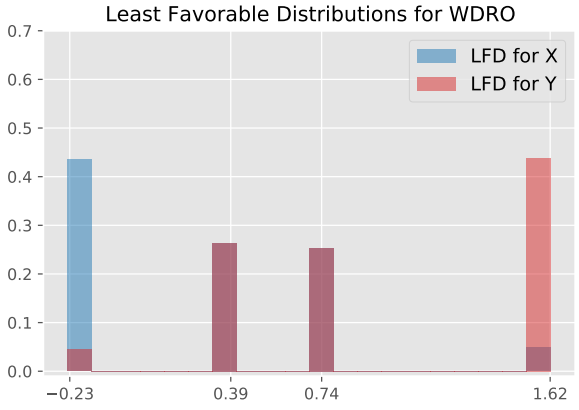
- $m = 10000$, $n = 1000$, takes 3.5 seconds on personal computer

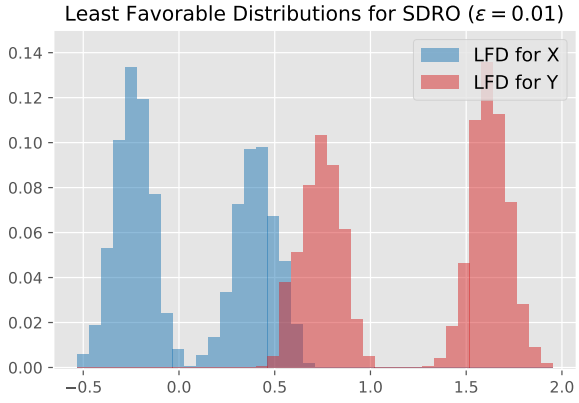
Thank discussions with A. Nemirovski, X. Cheng.

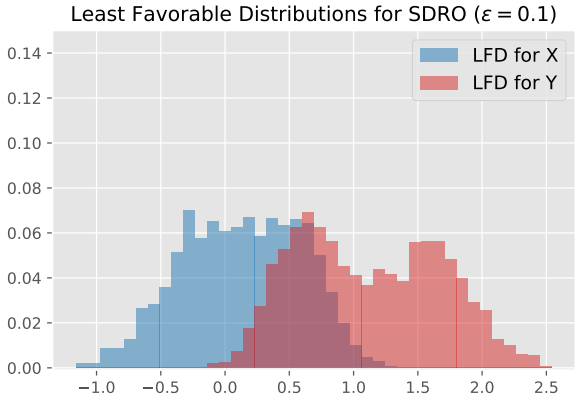
Example of LFDs

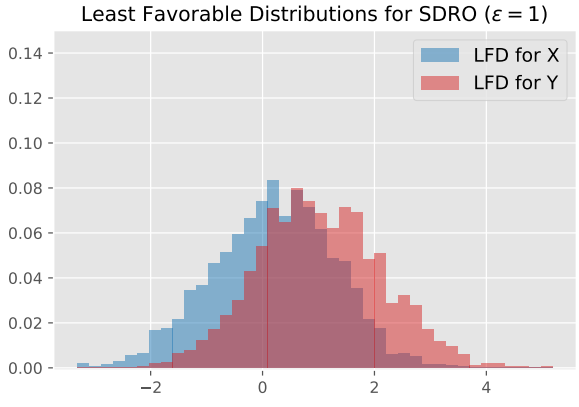
- Hypothesis testing based on 4 training samples:







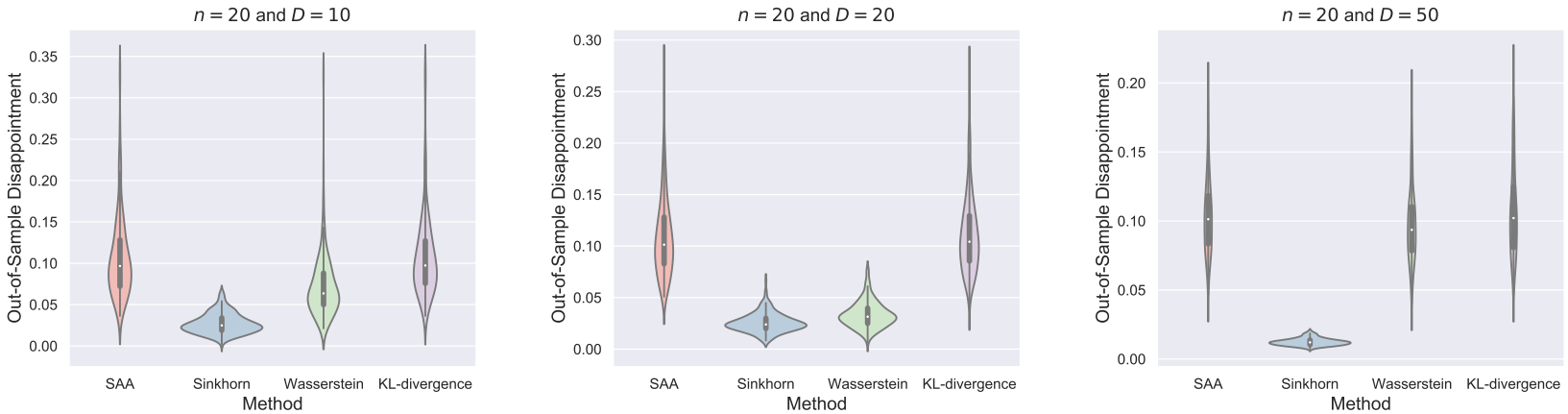




Example: Robust portfolio optimization

$$\min_x \quad \mathbb{E}_{P_*}\left[-x^T z\right] + \varrho \cdot P_*\text{-CVaR}_\alpha(-x^T z)$$
$$\text{s.t.} \quad 1^T x = 1$$

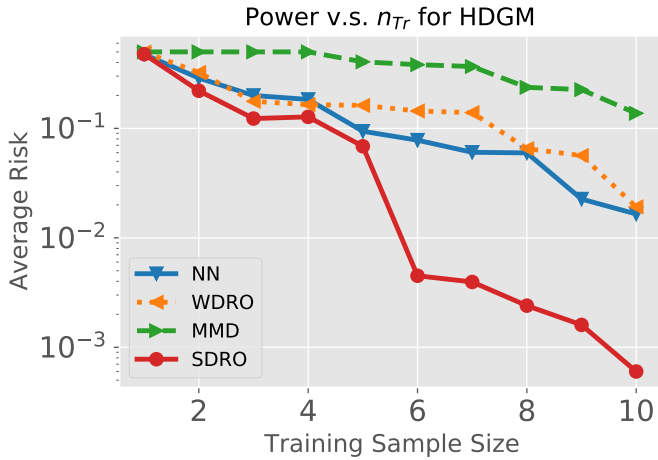
$z_i = \mathcal{N}(0, 0.02) + \mathcal{N}(0.03i, 0.025i),\ i = 1, \ldots, D$
Dimension $D \in \{10, 20, 50\}$ with sample size $n = 20$
$\alpha = 0.2,\ \varrho = 10$

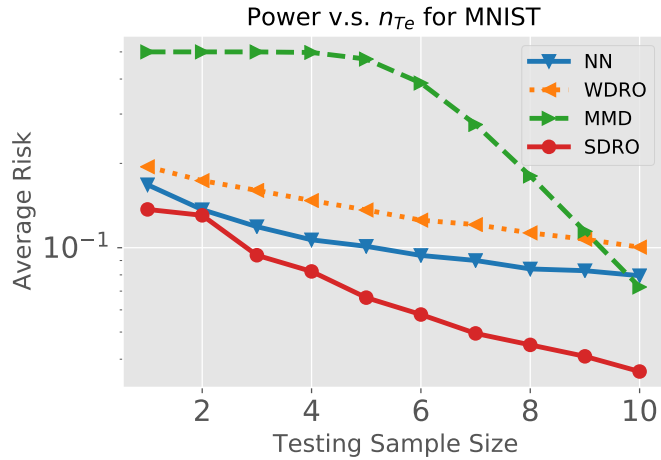


Smaller risk and smaller variance achieved by Sinkhorn DRO

Sinkhorn robust test: Performance comparison

- Example 1: Gaussian mixture (HDGM), data dimension 100
- Example 2: MNIST dataset for two digits
- Compare:
  neural network (NN) classify-to-test, Wasserstein/MMD/Sinkhorn robust tests
- Risk: Type I + Type II errors





Debiasing for Sinkhorn

- To define property divergence, need de-bias (Peyré, Cuturi 2020)

$$\tilde{S}_\epsilon(P,Q) := S_\epsilon(P,Q) - \frac{1}{2}S_\epsilon(P,P) - \frac{1}{2}S_\epsilon(Q,Q)$$

$\tilde{S}_\epsilon(P,P) = 0$

- Proposition (Feydy et al. 2022, Cheng, X. 2022): De-biased Sinkhorn divergence $\tilde{S}_\epsilon(P,Q)$ is convex in P for fixed Q.

- DRO induced by de-biased sinkhorn thus enjoys the properties we had before.

Feydy et al. Interpolating between optimal transport and mmd using sinkhorn divergences. AIStats, 2019

Debiasd Sinkhorn ensures "monotonicity"

- Compute $S_\epsilon(p, q)$ and $\tilde{S}_\epsilon(p, q)$ where $q = (1-t)p + tq_1$, $t \in [0, 1]$
- p and q_1: discrete distribution on 200 points

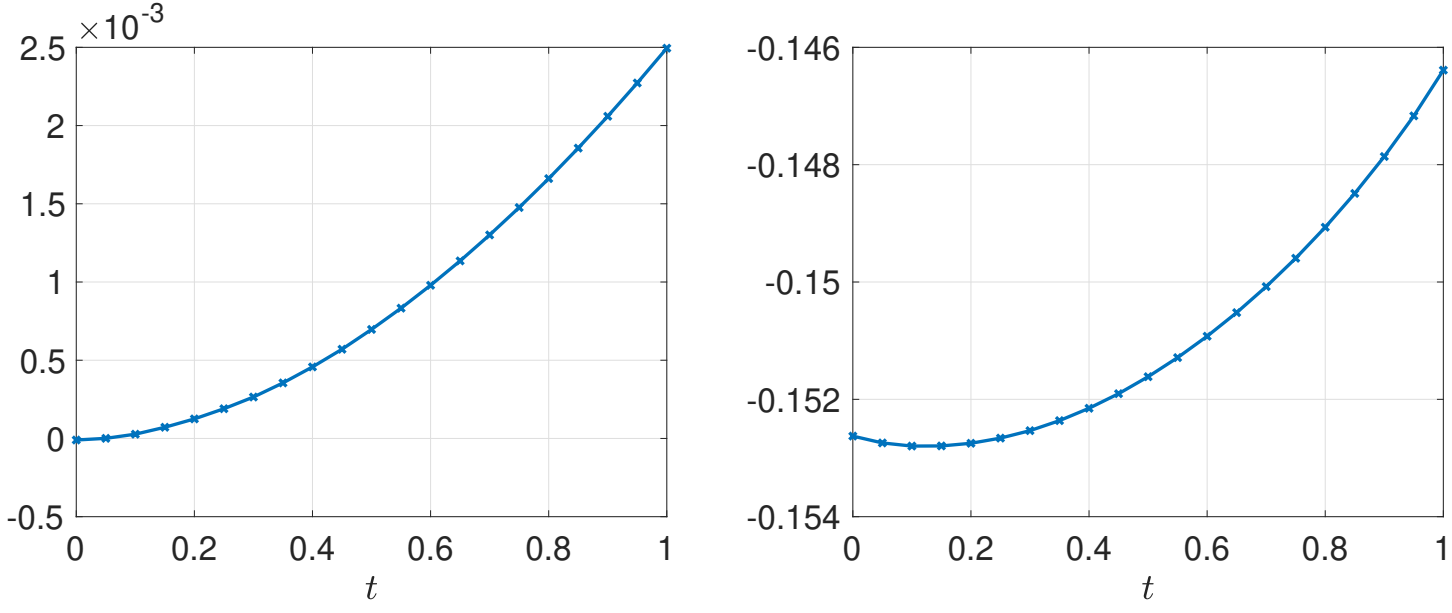


Figure: Left: With de-bias; Right: Without de-bias

Numerical example: DRO test with Debiased Sinkhorn

- Recall Φ is the risk of a detector ϕ:

$$\min_{\phi} \sup_{\tilde{S}_\epsilon(\hat{P}_k, P_k) \le \theta_k, k=1,2} \Phi(\phi; P_1, P_2)$$

  Debiased Sinkhorn

$$\tilde{S}_\epsilon(P, Q) := S_\epsilon(P, Q) - \frac{1}{2} S_\epsilon(P, P) - \frac{1}{2} S_\epsilon(Q, Q)$$

- In the following example, let $\theta_1 = \theta_2 = r$

Numerical results

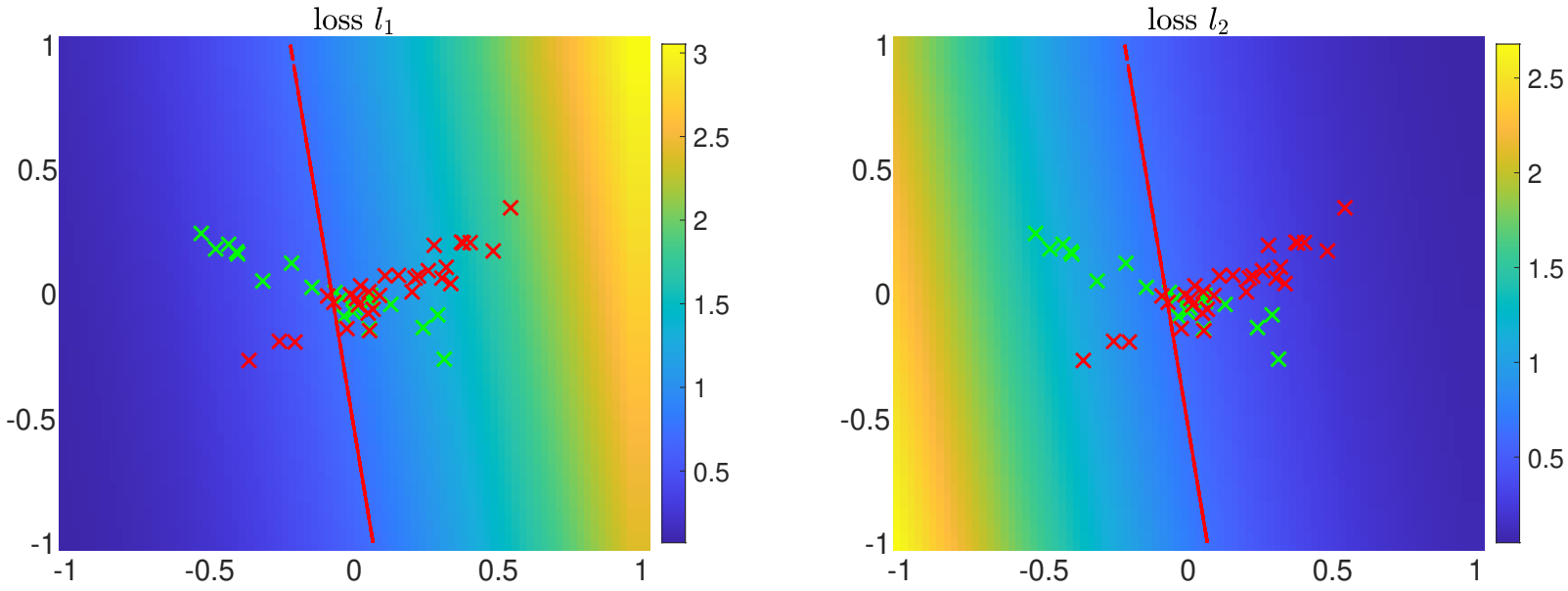


Figure: Loss of P_1 and P_2. Two data sets X_1 (green cross) and X_2 (red cross) having 20 and 30 points in $\mathbb{R}^2$

LFDs P_1 and P_2

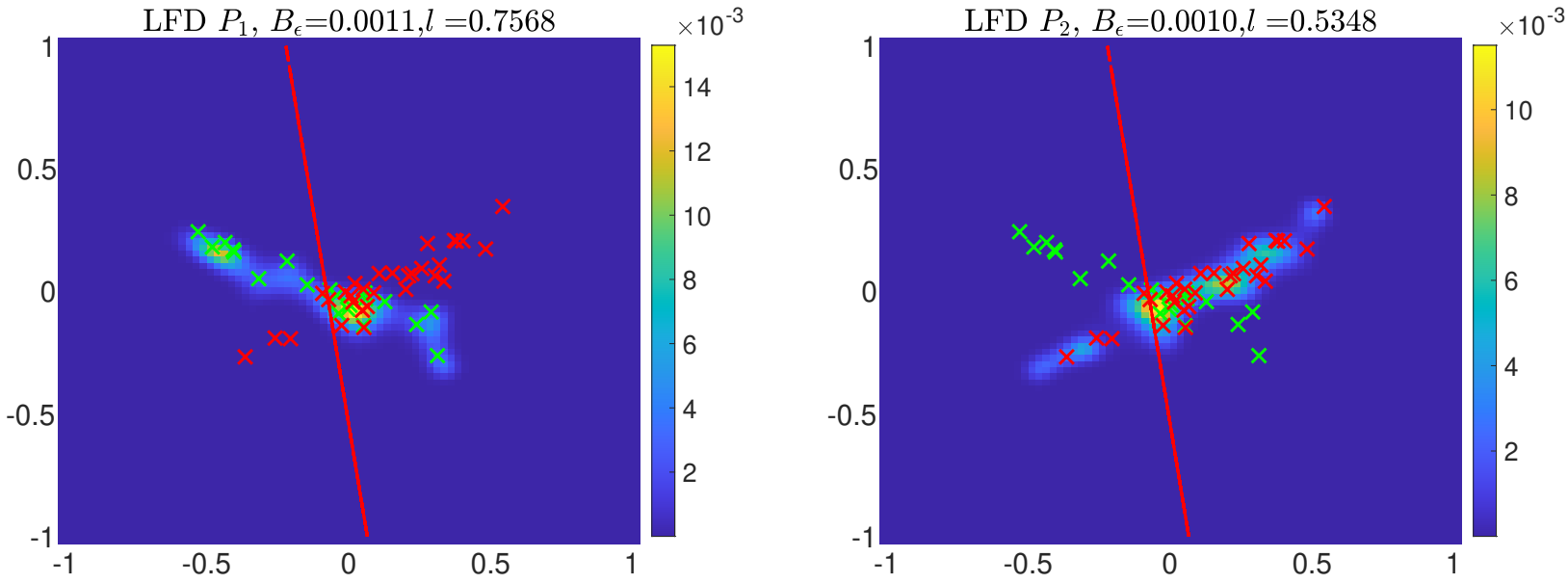


Figure: LFD illustrated as color field. $r = 0.001$.

Increase radius r, allow more perturbation in P_k.

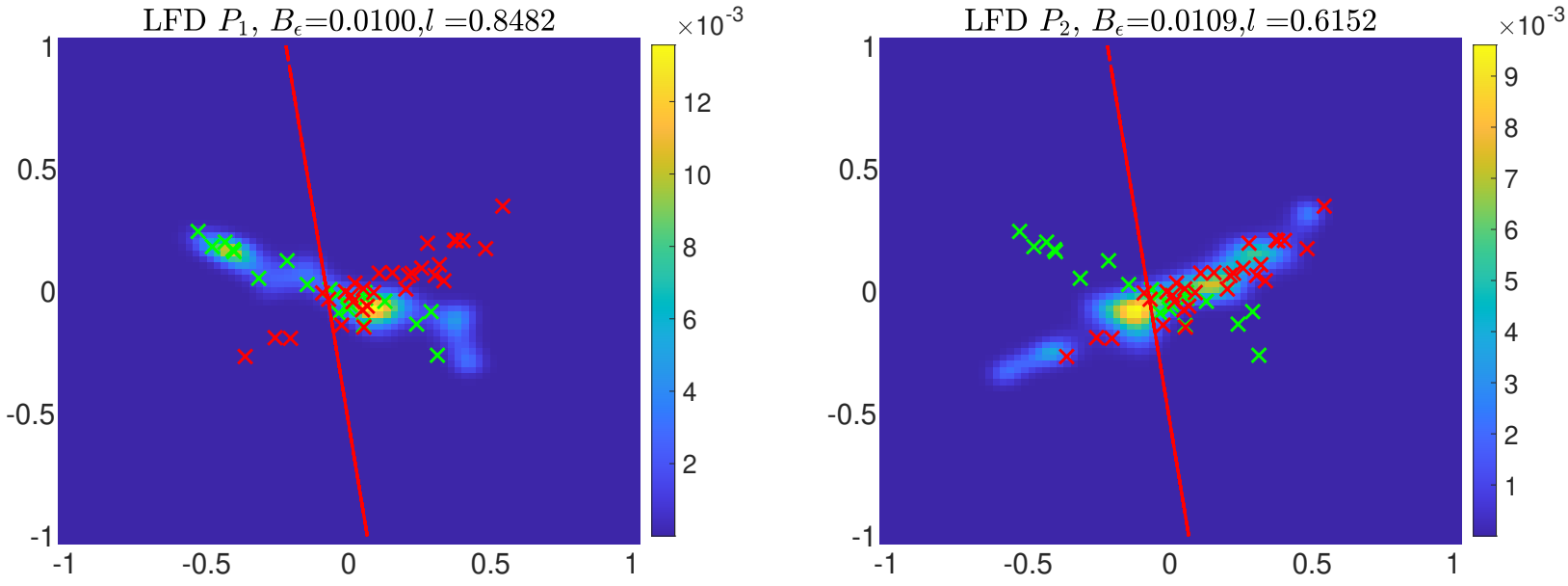


Figure: LFD illustrated as color field. $r = 0.01$.

Increase radius r, allow more perturbation in P_k.

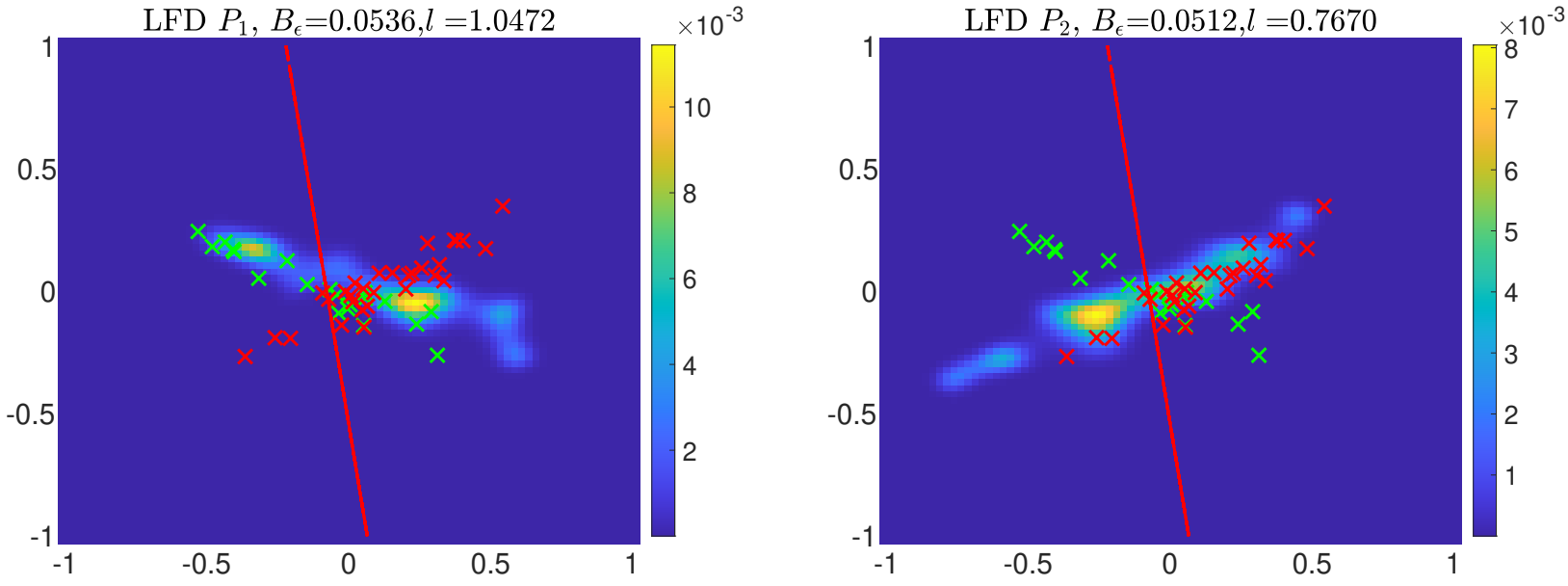


Figure: LFD illustrated as color field. $r = 0.05$.

Increase radius r, allow more perturbation in P_k.

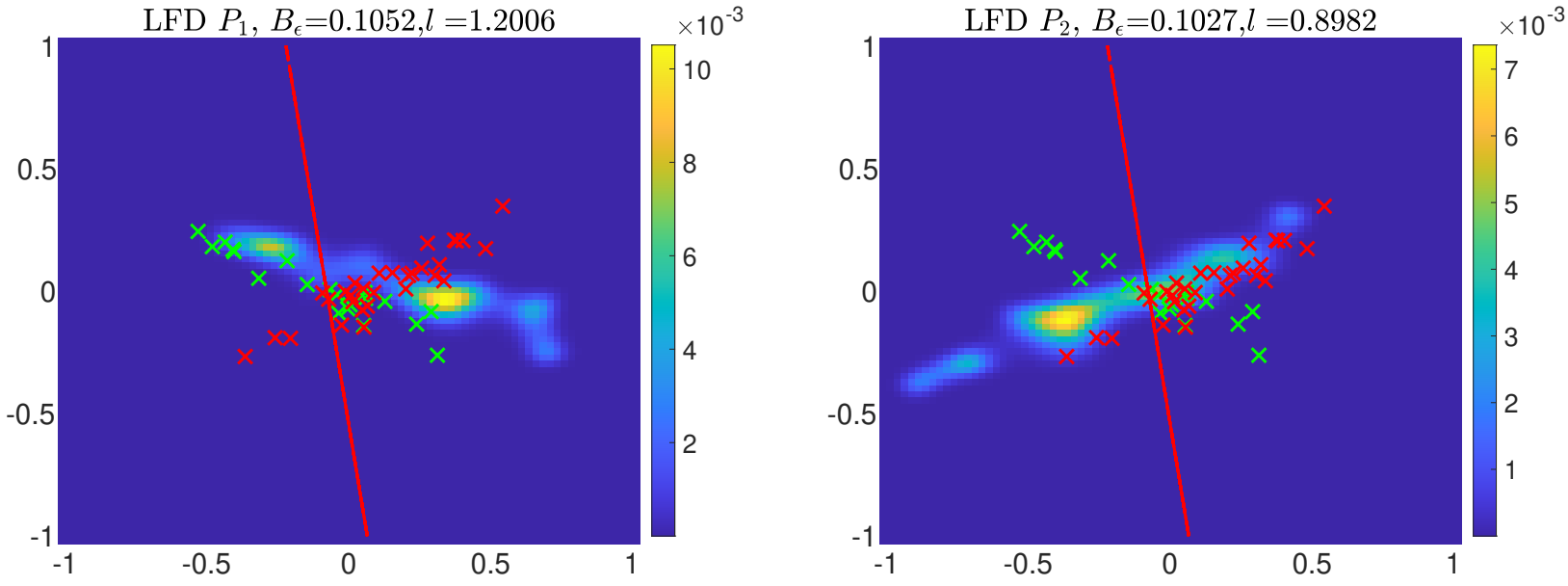


Figure: LFD illustrated as color field. $r = 0.1$.

Agenda

- Background
- Distributionally robust test: Wasserstein sets
- Distributionally robust test: Sinkhorn sets
- Extensions: classification, online detection $\leftarrow$
  (Joint work with Yang Cao, Shixiang Zhu, Liyan Xie, Rui Gao)

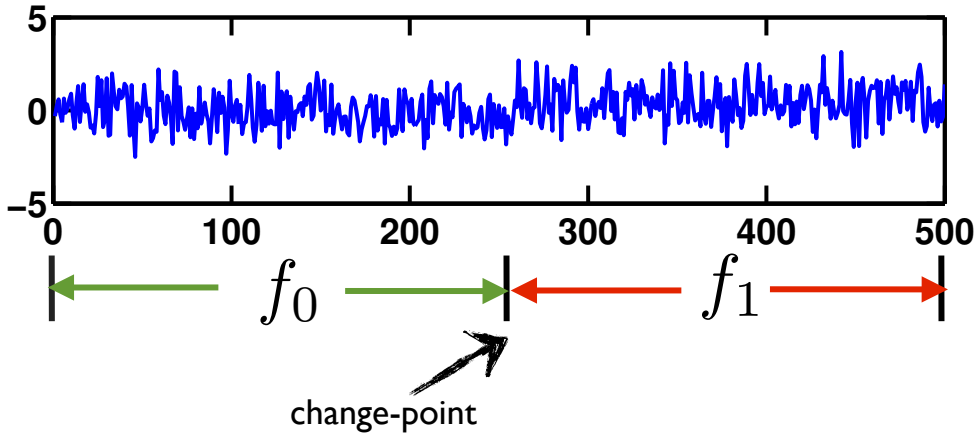
Sequential change-point detection

- Detecting a change that alters the data distribution

$$\begin{aligned} \mathsf{H}_0: &\quad x_t \sim f_0, \quad t = 1, 2, \ldots \\ \mathsf{H}_1: &\quad x_t \sim f_0, \quad t = 1, 2, \ldots, \kappa \\ &\quad x_t \sim f_1, \quad t = \kappa + 1, \ldots \end{aligned}$$

  unknown change-point location $\kappa > 0$

- Goal: detect change-point as quickly as possible



Robust CUSUM

(Cao, X. 2017)

- Use least favorable parameters solved from optimization problem to define a CUSUM procedure

$$\ell_t = \max\{0, \ell_{t-1} + \log \frac{f_{\theta_1^*}(x_t)}{f_{\theta_0^*}(x_t)}\}$$

- Detection procedure

$$T = \inf\{t : \ell_t > b\}$$

- ARL: $\mathbb{E}_\infty[T_1] \geq \gamma$ as long as $b \geq \log\gamma + \log\frac{\epsilon^*}{1-\epsilon^*}$,
- EDD $\leq \frac{b}{1-\epsilon^*}(1 + o(1))$
- Recall $1 - \epsilon^*$ Hellinger distance between two Gaussians

Example: Gaussian mean uncertainty

- For a Gaussian case, the optimization problem is explicit

$$\begin{aligned} \mathsf{H}_0: &\quad X \sim \mathcal{N}(\mu_0, \Sigma), \quad \mu_0 \in \mathcal{P}_0 \\ \mathsf{H}_1: &\quad X \sim \mathcal{N}(\mu_1, \Sigma), \quad \mu_1 \in \mathcal{P}_1 \end{aligned}$$

- Optimal detector $\phi^* = a^\intercal \omega - b$

$$a = \frac{1}{2}\Sigma^{-1}(\mu_0^* - \mu_1^*), \quad b = \frac{1}{2}a^\intercal \Sigma^{-1}(\mu_0^* + \mu_1^*)$$

$$(\mu_0^*, \mu_1^*) = \arg \min_{\mu_0 \in \mathcal{P}_0, \mu_1 \in \mathcal{P}_1} (\mu_0 - \mu_1)^\intercal \Sigma^{-1} (\mu_0 - \mu_1)$$

- Optimal risk

$$\epsilon^* = e^{-\frac{1}{8}(\mu_0^* - \mu_1^*)\Sigma^{-1}(\mu_0^* - \mu_1^*)}$$

Insights

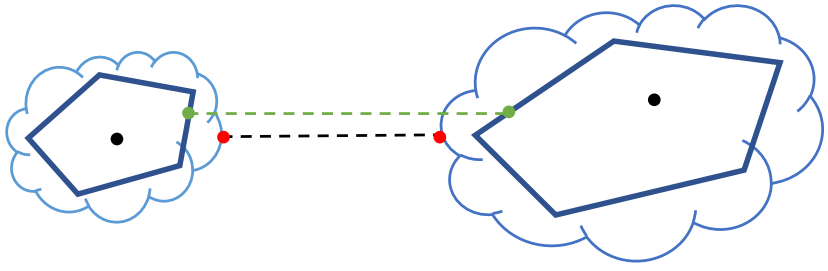
- Classic lower bound (Lorden 1971)

$$\text{EDD} \gtrsim \frac{b}{\text{KL}(f_1^* || f_0^*)}$$

- Our result says

$$\text{EDD} \lesssim \frac{b}{2(1-\epsilon^*)}$$

$\text{KL}(f_1^* || f_0^*) > 2 \cdot \text{Hellinger}$



True distributions f_0^* and f_1^*

Train neural networks to detect changes?

- Train a neural network by gradient descent
- Use *average difference in training loss* as test statistic

$$\hat{T}_{\text{net}} = \frac{1}{t}(\hat{L}(\theta_t) - \hat{L}(\theta_0)), t = 1, 2, \ldots$$

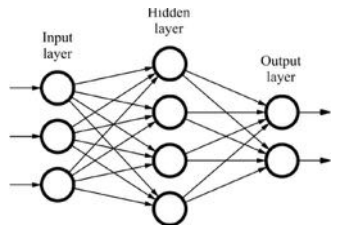
- Linear training loss

$$\hat{L}(\theta) = -\frac{1}{n_X}\sum_{i=1}^{n_X} f_\theta(x_i) + \frac{1}{n_Y}\sum_{i=1}^{n_Y} f_\theta(y_i)$$

f_θ: neural network

$x_i, i = 1, \ldots, n_X$

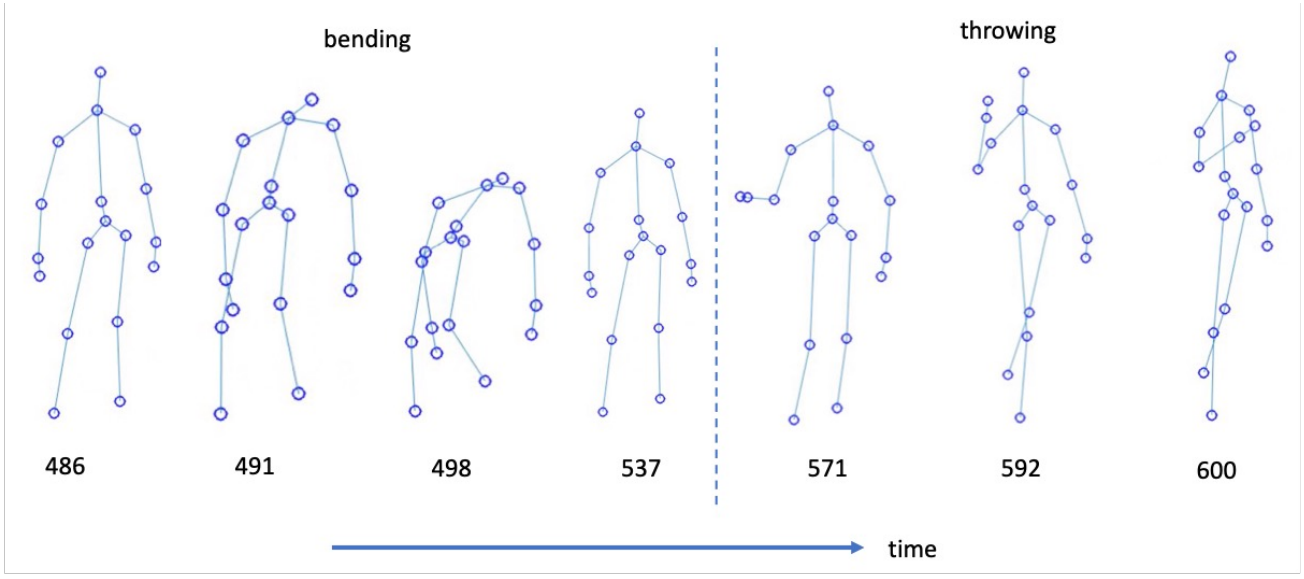
$y_i, i = 1, \ldots, n_Y$

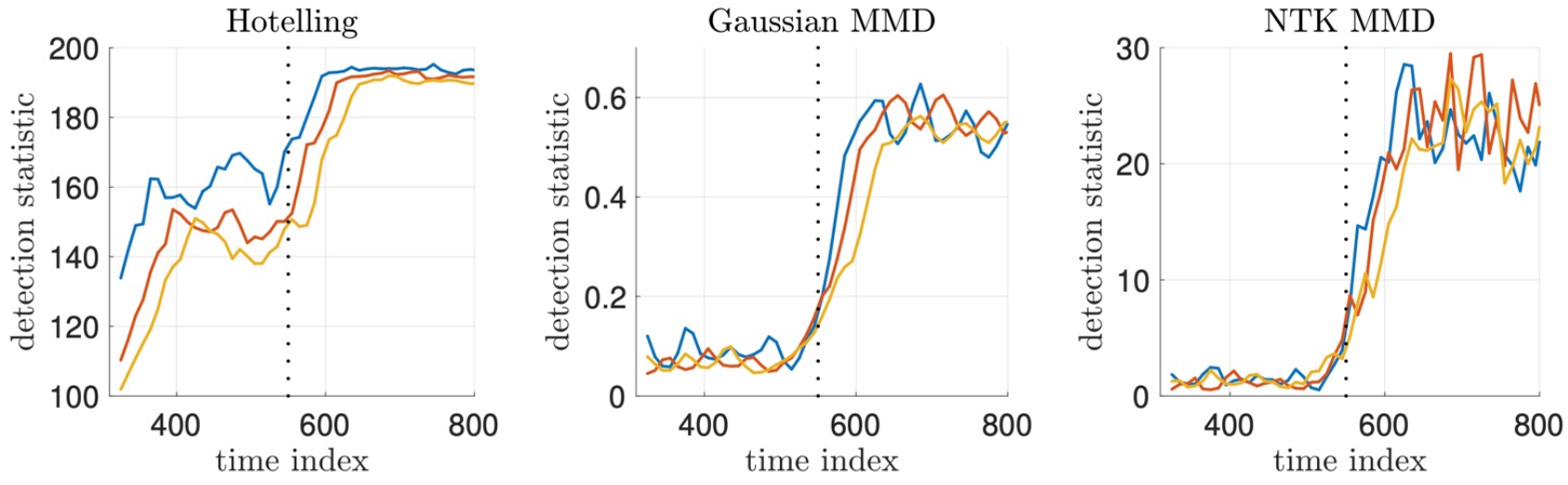


Test statistic

$$\hat{T}_{\text{net}}(t) := \frac{1}{t}\left(\hat{L}(\theta(t)) - \hat{L}(\theta(0))\right)$$

Real-data example: Human activity change detection





NTK-MMD

- Lazy training regime of overparameterized networks by gradient descent (Jacot, Gabriel, Hongler 18) (Chizat, Oyallon, Bach 19)
- Main technique

  Average difference in training loss =
  kernel MMD with neural-tangent-kernel

- Guarantee can be obtained by characterizing linearized training dynamics

$$\hat{T}_{\text{net}} \approx \iint K_0(x, x')(\hat{p}(x) - \hat{q}(x))(\hat{p}(x') - \hat{q}(x'))dxdx'$$

$$K_0(x, x') = \langle \nabla_\theta f(x, \theta_0), \nabla_\theta f(x', \theta_0) \rangle$$

- Type-I and Type-II errors guarantee by relating to kernels

Classification and hypothesis test

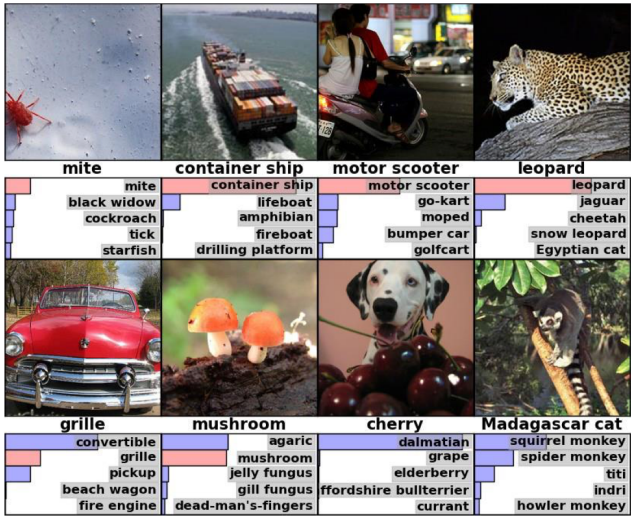
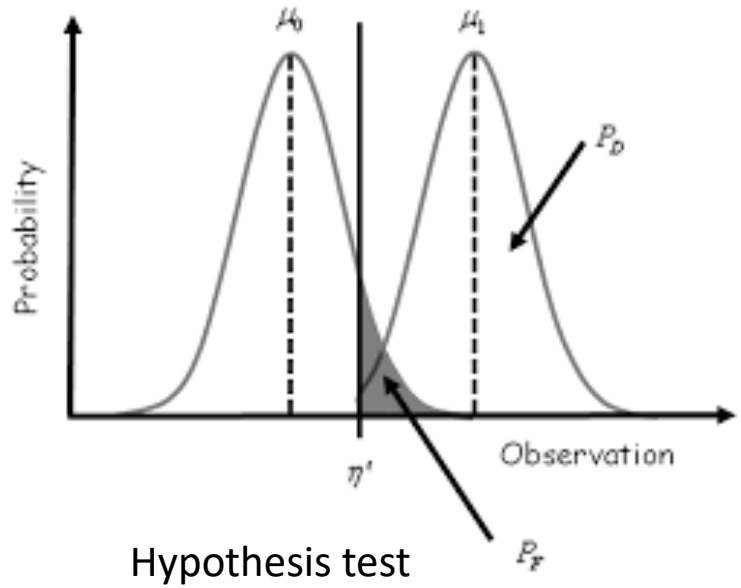


Image classification



Hypothesis test

- Classical statistical inference asks a question:
  Given observation ω, does it come from distributions p_1 or p_2?
- Classification typically considers more than 2 classes

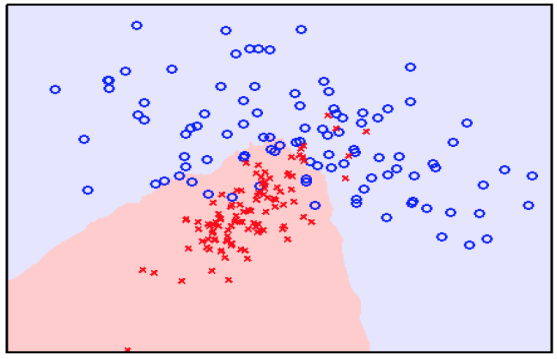
Vanilla k-nearest neighbors

- Classification with M classes
- Given n training samples: $(x_1, y_1), \ldots, (x_n, y_n)$
- Classify new test sample x using k-nn

(majority rule)

$$\hat{y}(x) = \arg \max_{m=1,\ldots,M} p_m(x)$$

$$p_m(x) = \sum_{i \in \mathcal{S}(x)} \frac{1}{k} \mathbb{I}\{y_i = m\}$$



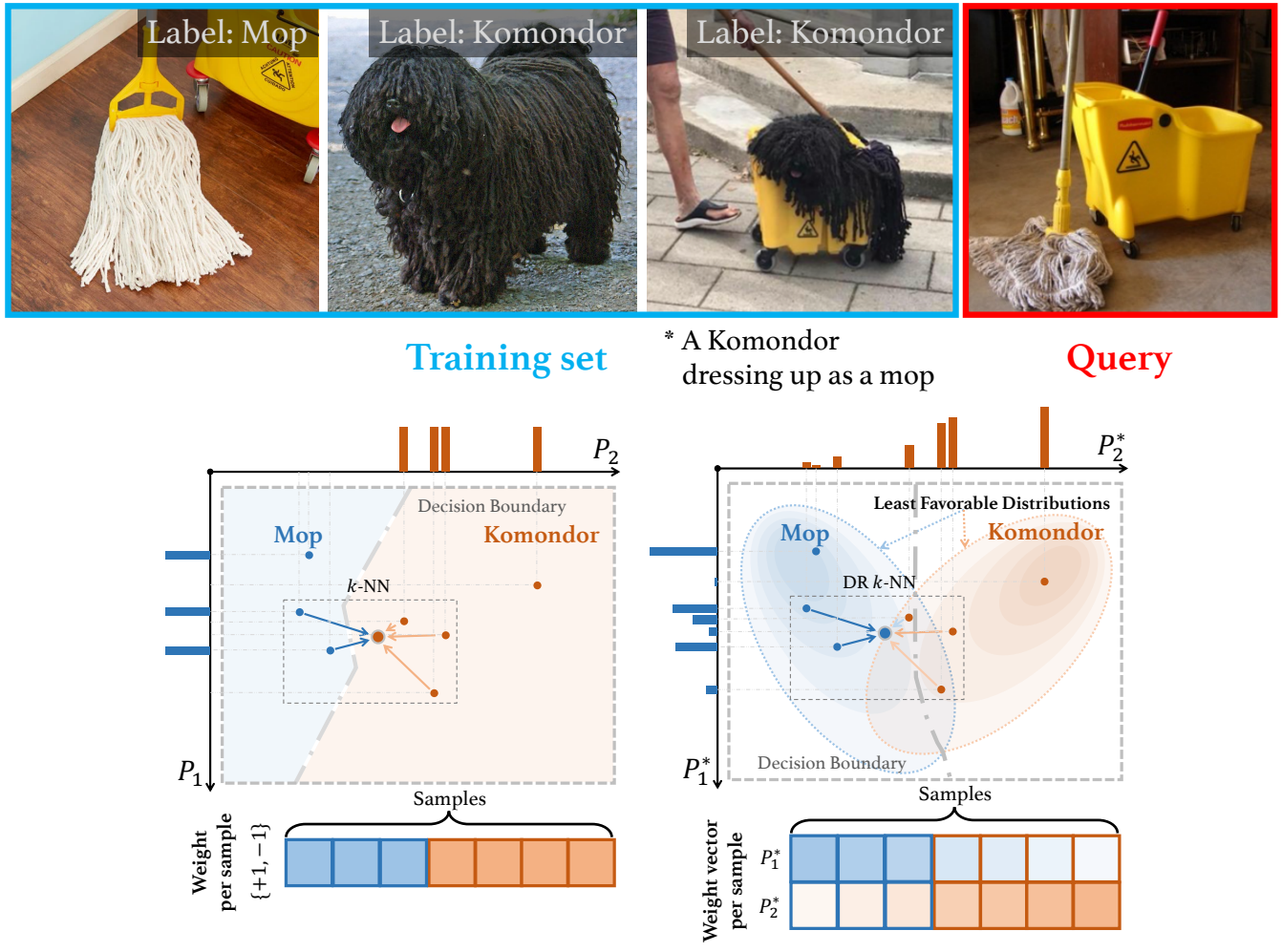
K = 25

$\mathcal{S}(x)$: nearest neighbor set, k training points with smallest $c(x, x_i)$

- Assign **equal** weights to all samples.

Distributionally robust k-nearest neighbor

Some samples may be more important than others.



Vanilla k-NN `DR. knn`

Distributionally-robust k nearest neighbor

- Assign **weights** to each training sample for each class

$$p_m(x) = \sum_{i \in \mathcal{S}(x)} w_m(x, x_i)$$

- Decision $\hat{y}(x) = \arg\max_{m=1}^{M} p_m(x)$
- Vanilla k-nn

$$w_m(x, x_i) = \mathbb{I}\{y_i = m\}$$

- Distance based weighted k-nn

$$w_m(x, x_i) = \frac{1}{c(x, x_i)} \mathbb{I}\{y_i = m\}$$

- Optimal weights: minimize asymptotic errors (Samworth 2012)

Distributionally robust k-nearest neighbor

- View weighted k-nn as a **random** decision function
- $\phi(x)$ decides class m w.p. $\pi_m(x)$, $\sum_{m=1}^{M} \pi_m(x) = 1$

$$\text{(Risk) } \Phi(\phi; P_1, \ldots, P_M) = \sum_{i=1}^{M} \mathbb{E}_{x \sim P_m}[1 - \pi_m(x)]$$

- Distributionally robust knn: $\pi_m^{(w)}(x) \propto p_m(x)$
- Find the optimal weights for worst-case distributions for all classes within the **uncertainty sets**

$$\min_{w} \max_{P_m \in \mathcal{P}_m} \Phi(\pi^{(w)}; P_1, \ldots, P_M)$$

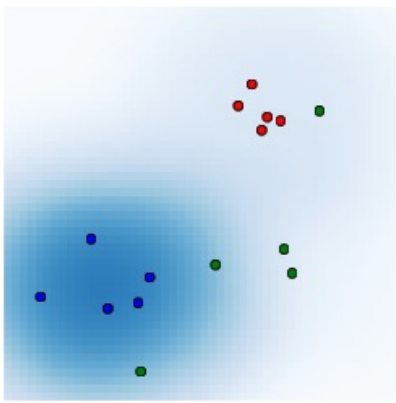
$\mathcal{P}_m$: Wasserstein uncertainty sets around empirical distributions

- **Convex** problem and can be solved efficiently

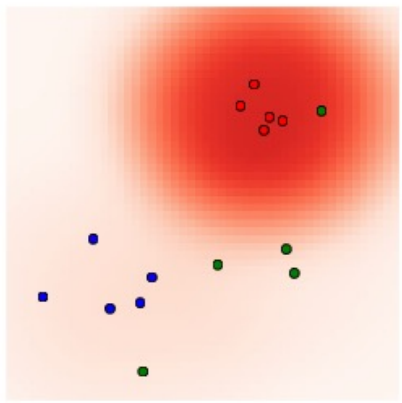
Example

- Example of the weights P_1^*, P_2^*, P_3^* using MNIST digit 4 (red), 6 (blue), 9 (green) and $k = 5$.

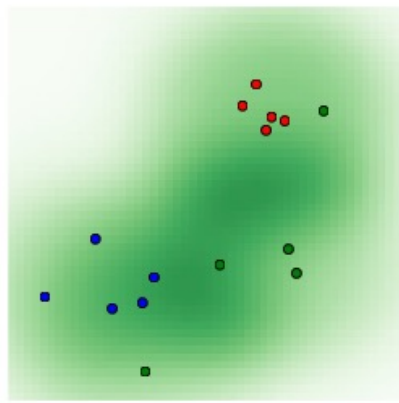
- Color represents kernel smoothed P_1^*, P_2^*, P_3^*



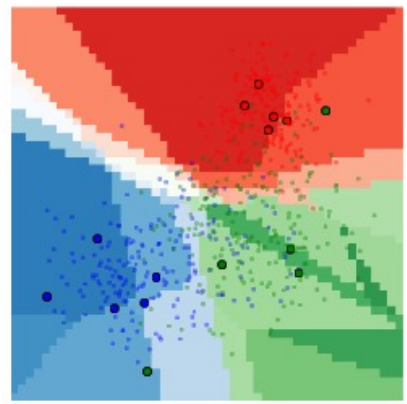
(a) P_1^*



(b) P_2^*



(c) P_3^*

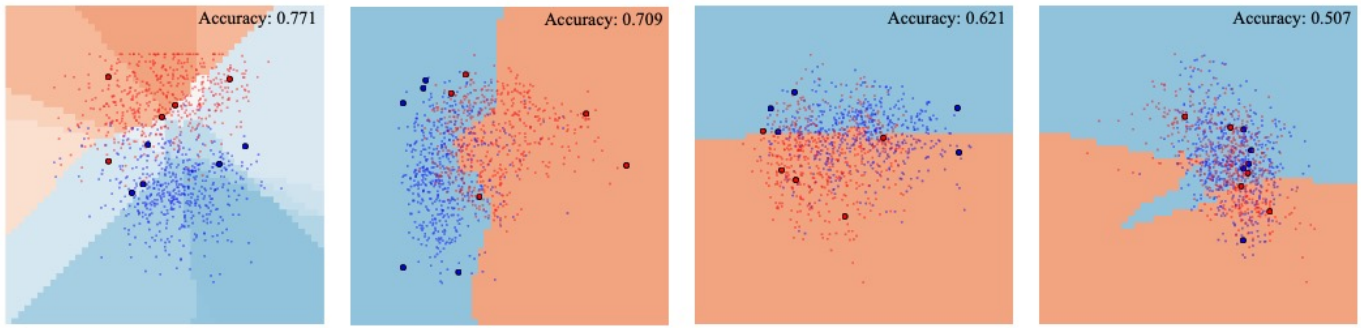


(d) π^*

Performance gain in few-shots learning

Table 1: Comparison of classification accuracy in the few-training-sample setting

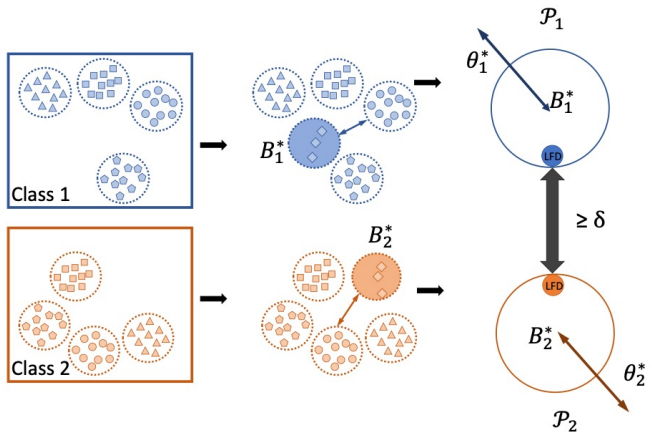
Methods	MNIST				*mini* ImageNet				CIFAR-10				Omniglot				Lung Cancer		COVID-19 CT	
	$M=2$		$M=5$		$M=2$		$M=5$		$M=2$		$M=5$		$M=2$		$M=5$		$M=3$		$M=2$	
	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=10$	$K=5$	$K=8$	$K=5$	$K=10$
PCA+k-NN	0.801	0.872	0.614	0.678	0.578	0.667	0.268	0.277	0.687	0.711	0.262	0.270	0.597	0.638	0.309	0.358	0.617	0.647	0.658	0.719
SVD+k-NN	0.749	0.790	0.524	0.567	0.587	0.675	0.268	0.283	0.680	0.701	0.259	0.266	0.591	0.618	0.305	0.413	0.624	0.648	0.646	0.715
NCA+k-NN	0.602	0.640	0.340	0.355	0.547	0.578	0.245	0.258	0.597	0.616	0.232	0.236	0.549	0.574	0.267	0.346	0.575	0.582	0.612	0.624
Matching Net	0.732	0.830	0.625	0.732	0.687	0.703	0.286	**0.360**	0.632	0.641	0.241	0.247	0.735	0.769	0.412	0.433	0.621	0.635	0.715	0.732
Prototypical Net	0.742	0.842	0.671	0.759	0.710	0.725	0.296	0.348	0.651	0.664	0.254	0.259	0.769	**0.836**	**0.448**	0.532	0.632	0.644	**0.729**	**0.744**
MetaOptNet	0.725	0.843	0.658	0.790	0.732	0.741	0.255	0.363	0.702	0.713	0.257	0.298	0.742	0.755	0.412	0.453	0.638	0.642	0.713	0.739
Feature embedding + k-NN	0.792	0.798	0.546	0.551	0.738	0.742	**0.490**	**0.486**	0.689	0.691	**0.492**	**0.494**	0.725	0.751	0.445	0.495	**0.664**	0.691	0.701	0.710
Kernel Smoothing	0.777	0.873	0.559	0.579	0.593	0.601	0.272	0.278	0.642	0.661	0.272	0.282	0.520	0.565	0.240	0.285	0.367	0.370	0.582	0.604
Truncated Dr.k-NN	**0.815**	**0.926**	**0.742**	**0.825**	**0.746**	**0.753**	0.295	0.340	**0.703**	**0.719**	0.297	0.305	**0.755**	0.825	0.425	**0.542**	0.652	**0.693**	0.722	0.741
Dr.k-NN	**0.838**	**0.959**	**0.746**	**0.831**	**0.752**	**0.786**	**0.306**	0.358	**0.707**	**0.728**	**0.309**	**0.311**	**0.765**	**0.850**	**0.465**	**0.580**	**0.667**	**0.704**	**0.734**	**0.752**



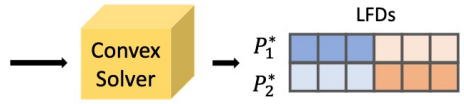
(a) `Dr.k-NN` (b) PCA + k-NN (c) SVD + k-NN (d) NCA.

Figure 5: A comparison of the learned feature spaces and the corresponding decision boundaries. There are 10 training samples from two categories of MNIST identified as large dots and 1,000 query samples identified as small dots. The color of dots shows their true categories. The color of the region shows the decisions made by corresponding methods.

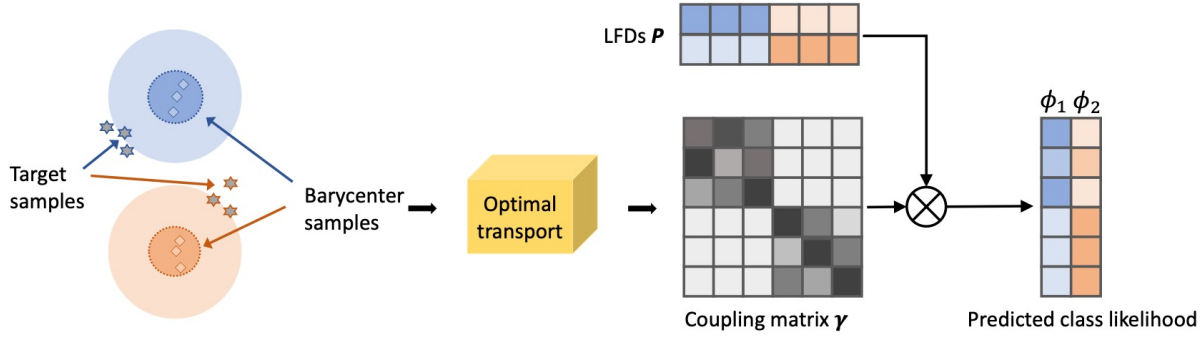
Distributionally-robust domain adaptation



(a) Construction of uncertainty sets.



(b) Distributionally robust optimization.

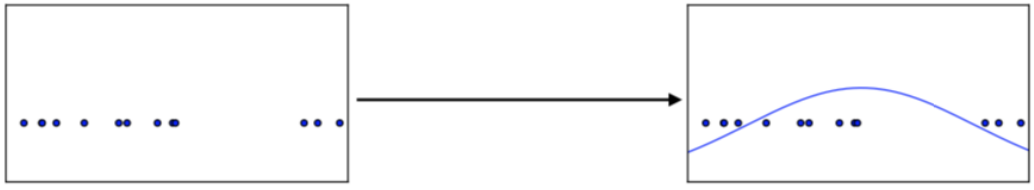


(c) Adaptive inference using optimal transport.

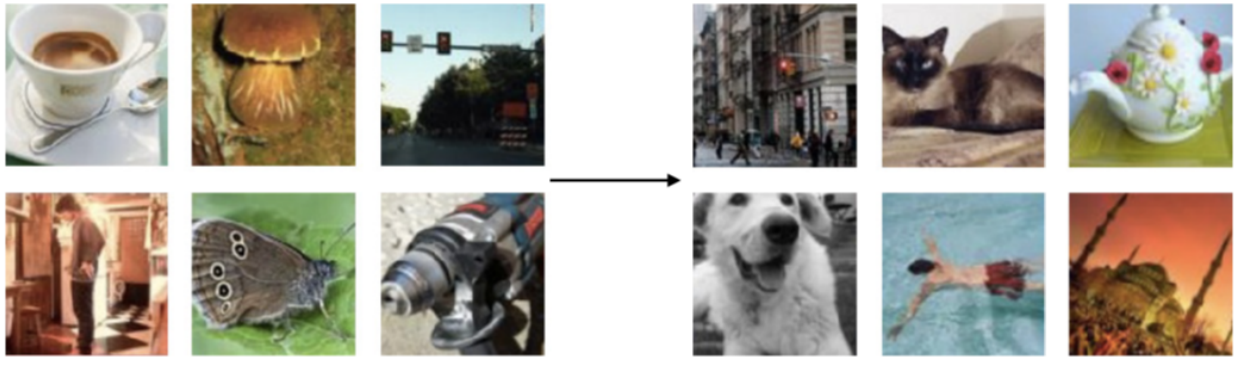
Generalizing to Unseen Domains with Wasserstein Distributional Robustness under Limited Source Knowledge, Wang, Xie, X., Huang, Li, 2022, arXiv:2207.04913.

Potential connection to GAN

- Generative adversarial networks (GAN)
  - Density estimation



  - Sample generation



Ian Goodfellow, 2014

GAN as minimax problem

- Minimax game

$$\min_{G} \max_{D} -\frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}} \log D(x) - \frac{1}{2}\mathbb{E}_{z} \log(1 - D(G(z;\theta)))$$

$D(\cdot)$: discriminator minimizes errors
$G(\cdot)$: generator maximizes the discriminator's error

- Loss resembles Jensen-Shannon divergence (heuristic)

- Use G^* generative models: ofter produce the good samples

Ian Goodfellow, 2016 NeurIPS tutorial

Minimax hypothesis test

- minimax test

$$\inf_{\phi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_1} \underbrace{\mathbb{E}_{P_1}[\mathbb{I}\{T(x) = 2\}]}_{\text{Type-I error}} + \underbrace{\mathbb{E}_{P_2}[1 - \mathbb{I}\{T(x) = 2\}]}_{\text{Type-II error}}.$$

- GAN

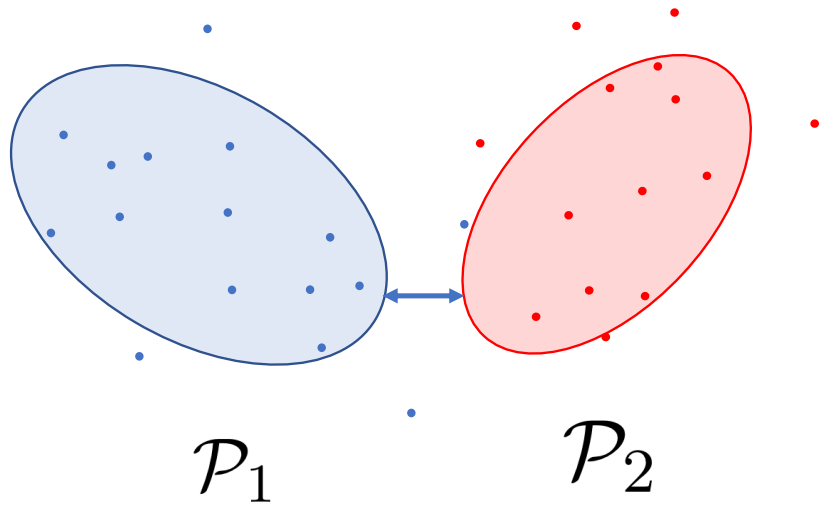
$$\min_{D} \max_{p_{\text{generator}}} \mathbb{E}_{p_{\text{data}}} \log D(x) + \mathbb{E}_{p_{\text{generator}}} \log(1 - D(x))$$

GAN	Hypothesis test
discriminator D	test ϕ
generator G	least favorable distribution P_1^*, P_2^*

Summary

Robust hypothesis test

$$\inf_{\phi} \sup_{P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2} \Phi(\phi; P_1, P_2)$$



- Distributionally robust hypothesis test
- Uncertainty sets: Wasserstein, sinkhorn
- Computational efficient convex reformulation
- Extension to classification, online detection, domain adaptation

Thanks NSF DMS-2134037, CAREER CCF-1650913.

References

- Minimax Robust Hypothesis Testing. Gul, Zoubir (2017)
- Robust sequential change-point detection by convex optimization. Cao, and X. ISIT. (2017).
- Hypothesis testing by convex optimization. Goldenshluger, Juditski, Nemirovski. (2015)
- Robust hypothesis testing with Wasserstein uncertainty sets. Gao, Xie, X. NeurIPS 2018 (Spotlight). arXiv:2105.14348 (2021)
- Sinkhorn distributionally robust optimization. Wang, Gao, X. arXiv:2109.11926 (2021)
- A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. Wang, X. ISIT. (2022)
- Distributionally robust k-nearest neighbors. Zhu, Xie, Zhang, Gao, and X. arXiv:2006.04004 (2020)
- Computational optimal transport. Peyré, Cuturi. (2020)
- Neural tangent kernel maximum mean discrepancy. Cheng, and X. Neurips. (2021)

Thank NSF DMS-2134037, 1938106, 1830210, CCF-1650913, CMMI-2015787, 1538746.