Kernel Landmark: An Empirical Statistical Approach to Detect Covariate Shift

Yuksel Karahan, Bilal Riaz, Austin J. Brockmeier

Electrical and Computer Engineering Department, University of Delaware, Newark DE

What is the goal?

Detect: Divergence between train and test

Identify: Specific examples for the discrepancy

Training and test input follow different distributions, but functional relation remains unchanged.

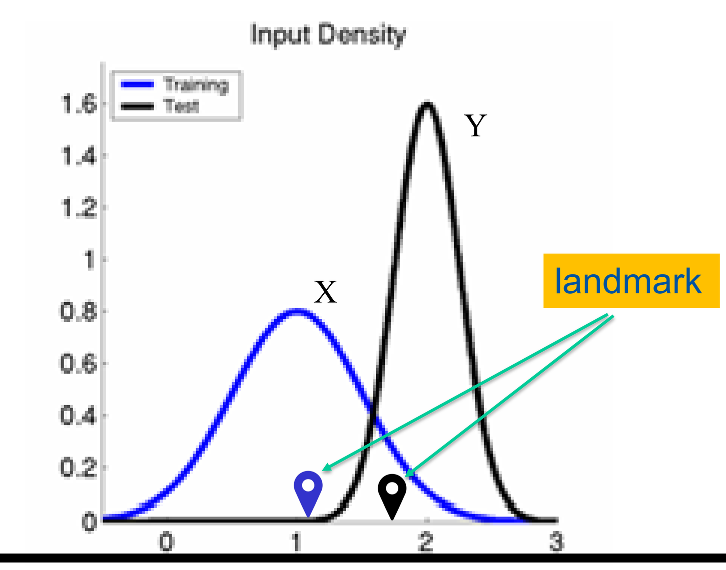


Image credit: http://www.ms.k.u-tokyo.ac.jp/sugi/

Two-Sample Tests using Kernel Divergences

- Probability measures are $\mu, \nu \in P(\mathcal{X})$
- The empirical measures are $\hat{\mu} = \sum_i^m \mu_i \delta_{x_i}$ and $\hat{\nu} = \sum_i^n \nu_i \delta_{y_i}$

Maximum mean discrepancy (MMD)

$$\mathrm{MMD}^{\mathcal{H}}(\mu,\nu) = \sup_{\omega\in\mathcal{F}} \mathbb{E}_{X\sim\mu, Y\sim\nu}[\langle \phi(X)-\phi(Y), \omega\rangle] = \sup_{\omega\in\mathcal{F}} \mathbb{E}[\omega(X)-\omega(Y)] = \|m_\mu - m_\nu\|_{\mathcal{H}}$$

The max-sliced kernel Wasserstein-2 (W2)

$$W_2^{\mathcal{H}_*}(\hat{\mu},\hat{\nu})^2 = \max_{\alpha\in\mathcal{A}} \min_{\mathbf{P}\in\mathcal{P}_{\hat{\mu},\hat{\nu}}} \left\{ \sum_{i,j} P_{ij} |\omega(x_i) - \omega(y_j)|^2 = \langle \mathbf{P}, (\mathbf{K}_{XZ}\alpha \mathbf{1}_n^\top - \mathbf{1}_m (\mathbf{K}_{YZ}\alpha)^\top)^{\circ 2} \rangle \right\}$$

$$= \max_{\alpha\in\mathcal{A}} \langle \boldsymbol{\mu}, (\mathbf{K}_{XZ}\alpha)^{\circ 2}\rangle + \langle \boldsymbol{\nu}, (\mathbf{K}_{YZ}\alpha)^{\circ 2}\rangle - 2 \max_{\mathbf{P}\in\mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}\mathbf{K}_{YZ}\alpha, \mathbf{K}_{XZ}\alpha\rangle$$

where $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{XZ} \\ \mathbf{K}_{YZ} \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}$ is the kernel matrix.

Minimax optimization: in which evaluation requires $O(N \log N)$

Proposed Methods : Kernel Landmarks

- **Landmark max-sliced kernel Wasserstein (L-W2)**

At most $l = 2N$ evaluations each requires $O(N \log N)$

i-th landmark

$$W_2^{\mathcal{H}_{L^*}}(\hat{\mu},\hat{\nu}) = \sqrt{\max_{i\in\{1,\dots,l\}} \frac{1}{N} \sum_j (\kappa(x_{R_i(j)}, z_i) - \kappa(y_{Q_i(j)}, z_i))^2}$$

Permutations based on i-th landmark

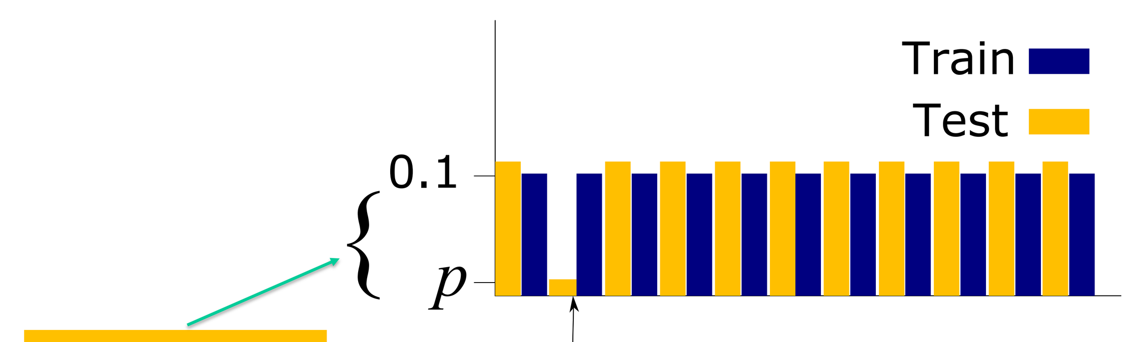
```
% K - kernel matrix where K(i,j) = kappa(Z{i}, Z{j})
% S - binary indicator for points in Z being from X
[val, i_star] = max(mean( (sort(K(:,S==1), 2) - sort(K(:,S==0), 2) ).^2, 2) );
div = sqrt(val);
```

- **Landmark max-sliced kernel Bures (L-Bures)**

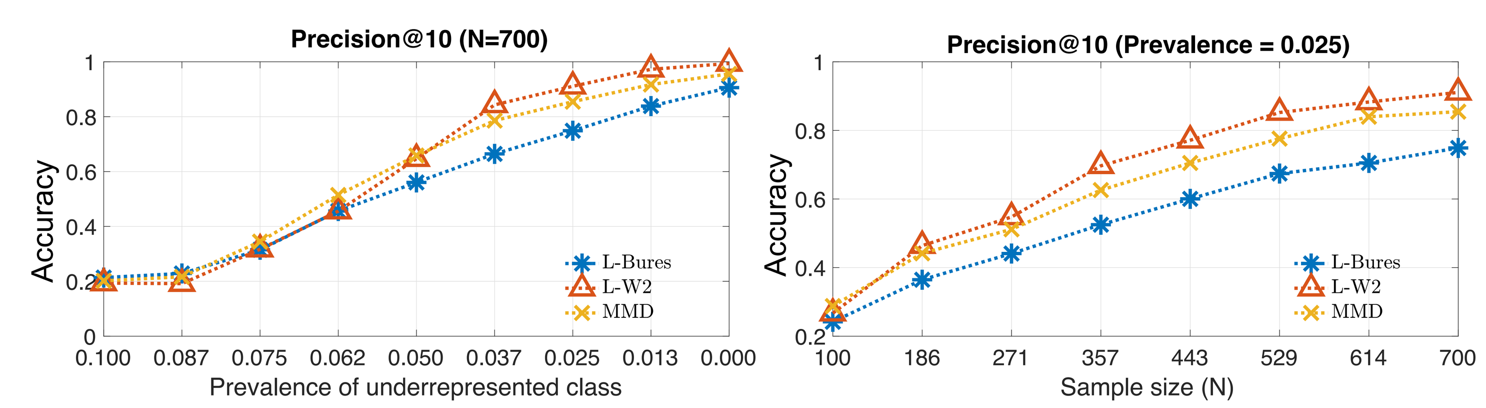
At most $l = 2N$ evaluations each requires $O(N)$

$$D_B^{\mathcal{H}_{L^*}}(\hat{\mu},\hat{\nu}) = \max_{i\in\{1,\dots,l\}} \left\{ \left| \frac{1}{\sqrt{m}} \|\mathbf{k}_{Xz_i}\|_2 - \frac{1}{\sqrt{n}} \|\mathbf{k}_{Yz_i}\|_2 \right| \right\}$$

Identify the Missing Class using Witness Function



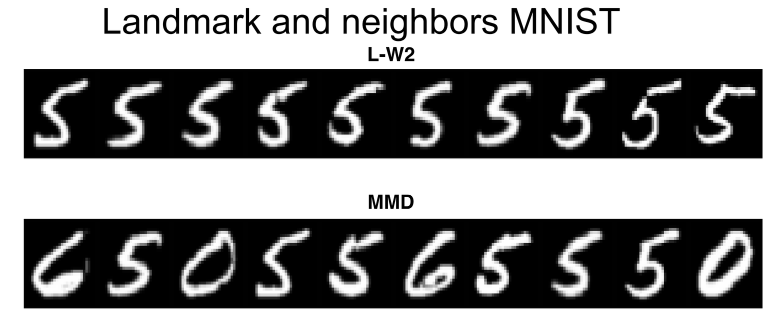
Precision of the witness function in detecting minority classes



Mean precision accuracy for digit "0" across 500 Monte Carlo iterations (MNIST).

Missing Digit: 5

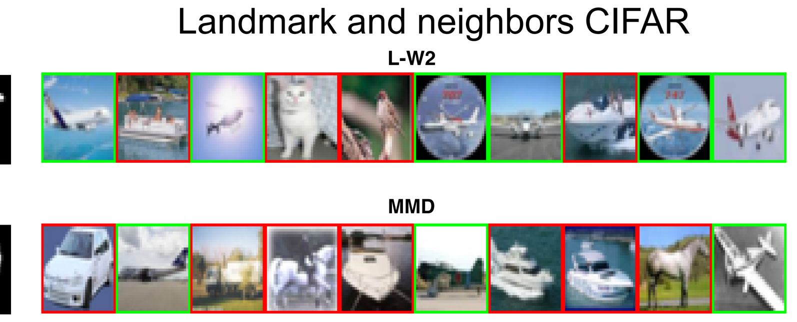
Landmark and neighbors MNIST



Maximal discrepancy points MNIST

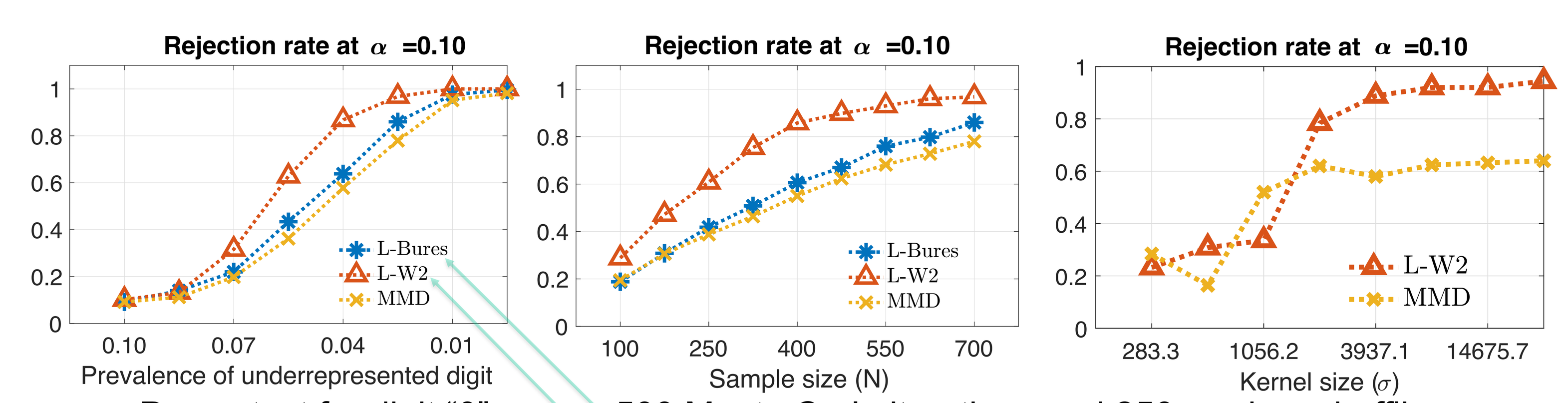
Missing Class: Airplane

Landmark and neighbors CIFAR



Maximal discrepancy points CIFAR

Statistical power test as a function of the class prevalence and sample size, and kernel sizes on MNIST dataset

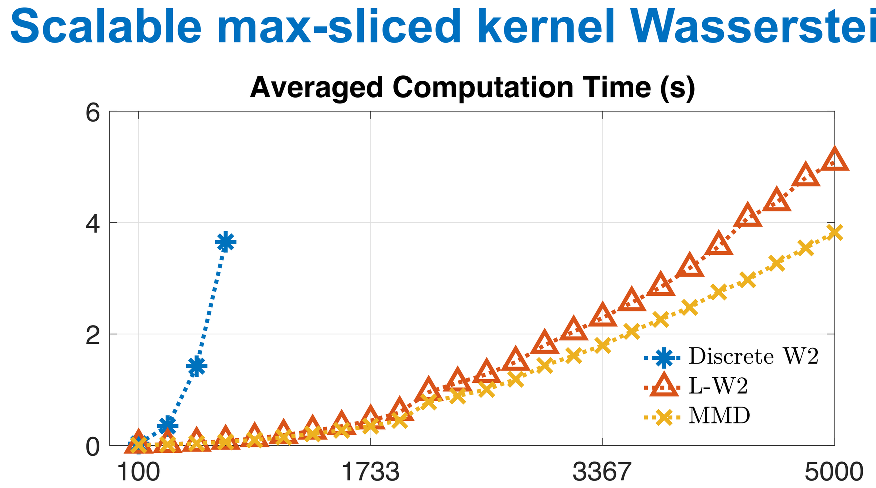


Power test for digit "0" across 500 Monte Carlo iterations and 250 random shuffling.

L-Bures: Landmark max-sliced kernel Bures

L-W2: Landmark max-sliced kernel Wasserstein

Scalable max-sliced kernel Wasserstein



The implementation of our experiment can be found by scanning QR-code above.

Detecting Data Changes using Different Learning Representations

No Reduction (NoRed): which is the raw data

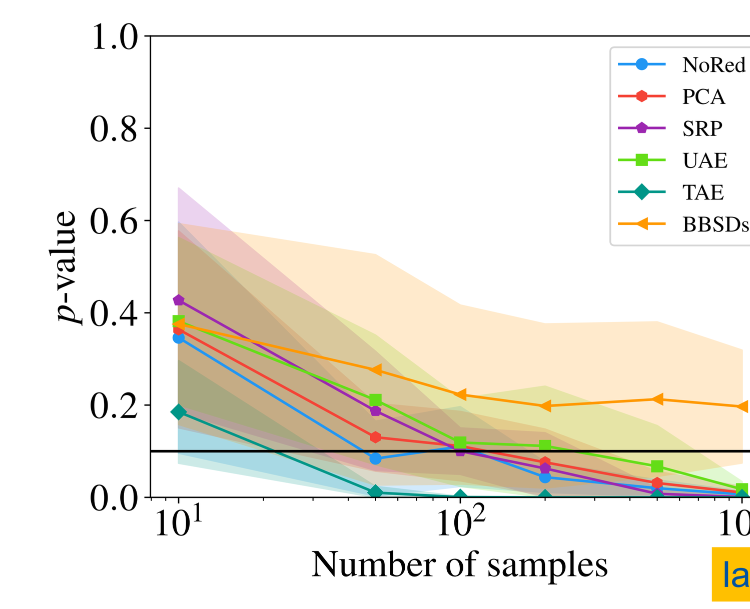
Principal Component Analysis (PCA): $\hat{X} = XR$

Sparse Random Projection (SRP): $\hat{X} = XR$

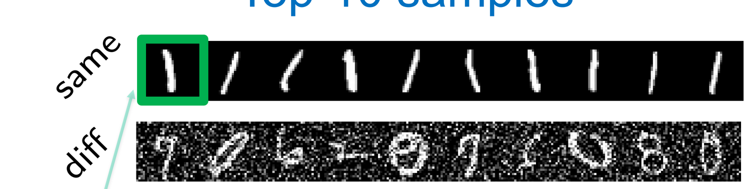
Autoencoders (AE): [Untrained (UAE) & Trained (TAE)] $h = \varphi(x)$

Black Box Shift Detection (BBSD): using softmax outputs

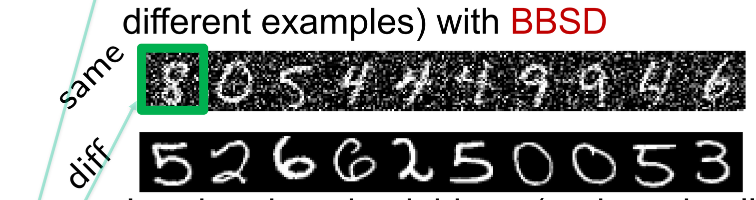
Perturbation: Gaussian Noise (SNR=0.4)



Top-10 samples



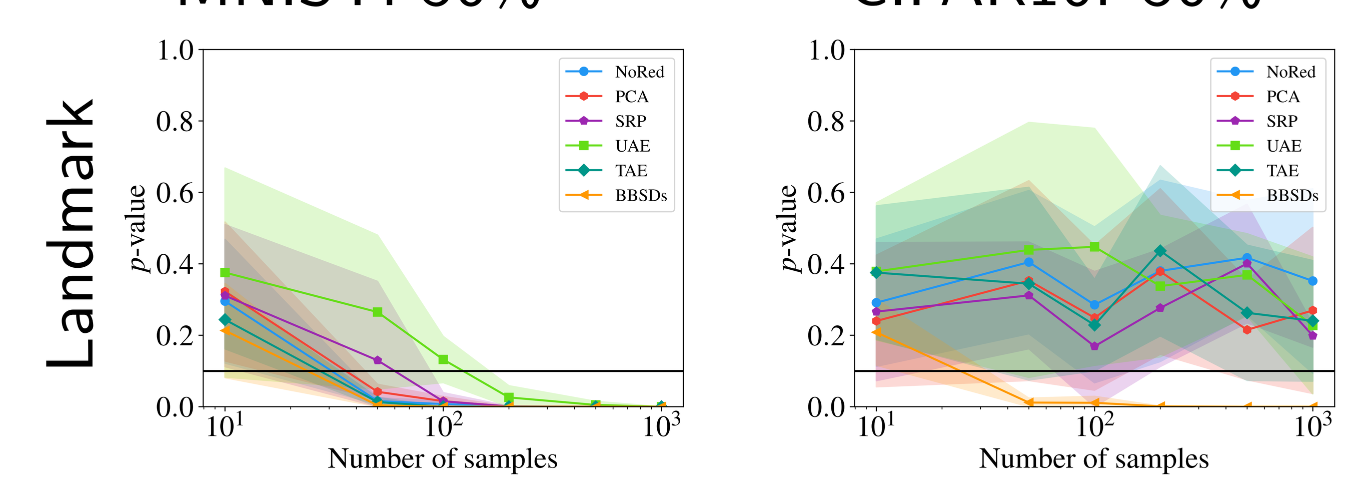
Landmark and neighbors (and maximally different examples) with BBSD



Landmark and neighbors (and maximally different examples) with TAE

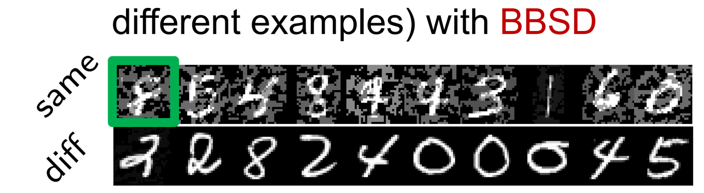
Perturbation: Adversarial Shift

MNIST: 50% — CIFAR10: 50%



Top-10 samples

Landmark and neighbors (and maximally different examples) with BBSD



Landmark and neighbors (and maximally different examples) with TAE

Top-10 samples

Landmark and neighbors (and maximally different examples) with BBSD

Landmark and neighbors (and maximally different examples) with TAE

Conclusion

- We have investigated max-slicing for the kernel-based Wasserstein distance to detect class-based covariate shift.
- Our approach evaluates the discrepancy between distributions.
- The proposed distance can be computed exactly and efficiently for the case of two samples.
- The preliminary results shows that the proposed method detects simple cases of covariate shift better than MMD.

Thanks to ONR (N00014-21-1-2300) for the support