Kernel Landmarks: An Empirical Statistical Approach to Detect Covariate Shift

Yuksel Karahan, University of Delaware
Bilal Riaz, University of Delaware
Austin J. Brockmeier, University of Delaware

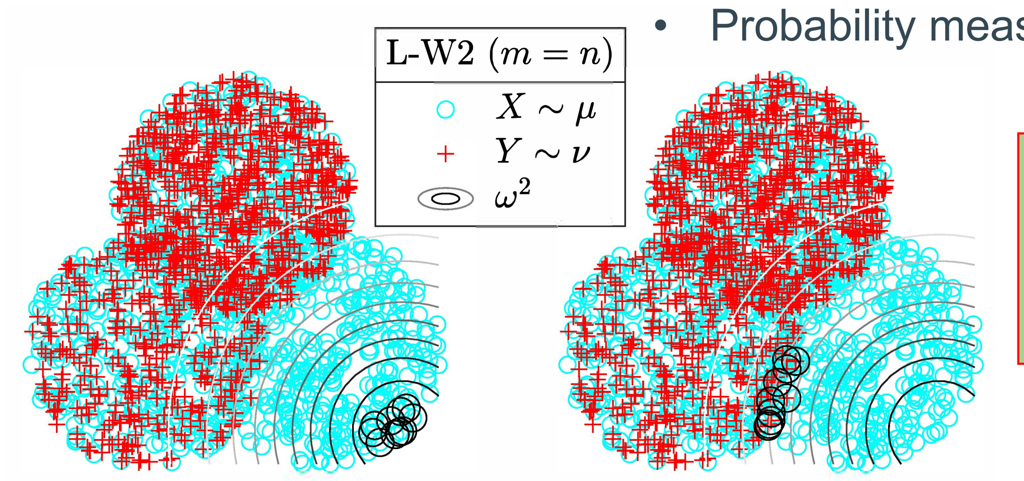
UNIVERSITY OF DELAWARE

NEURAL INFORMATION PROCESSING SYSTEMS

- **We propose an alternative solution to kernel max-slicing**
- **Each data point (landmark) defines a witness function**
- **The landmark which identifies the largest discrepancy between the distribution is chosen**
- **Our approach detects class-based covariate shift**
- **It identifies instances from minority class based on witness functions**
- **The landmark-based kernel max-slicing is much simpler to compute than the kernel max-slicing**

What is the goal?

Divergence measures for interpreting and minimizing discrepancies between data distributions.



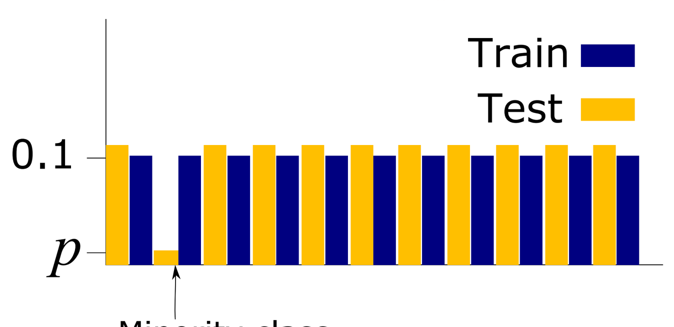
- Probability measures are $\mu, \nu \in P(\mathcal{X})$

Examine the top-10 examples (in terms of witness function) from each sample

Covariate shift: When the testing cases are not class-balanced

By localizing discrepancies

- **Detect:** Divergence between train and test
- **Identify:** Classes for witness's top-K training set examples



- p is the class prevalence

Two-sample Tests Using Kernel Divergences

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)

$$W_2^{\mathcal{H}_*}(\mu,\nu) = \sup_{\omega\in\mathcal{F}} W_2(\omega_\sharp\mu, \omega_\sharp\nu) = \sup_{\omega\in\mathcal{F}} \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\gamma}\left[|\omega(X)-\omega(Y)|^2\right]\right)^{\frac{1}{2}}$$

Empirical measures formed from two samples: $\hat{\mu} = \sum_i^m \mu_i\, \delta_{x_i}$ and $\hat{\nu} = \sum_i^n \nu_i\, \delta_{y_i}$

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

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

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.

Saddlepoint optimization problem: evaluation requires $\mathcal{O}(N\log N)$

Kernel Landmarks

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

$$W_2^{\mathcal{H}_{L^*}}(\mu,\nu) = \sup_{z\in\mathcal{X}} \inf_{\gamma\in\Gamma(\mu,\nu)} \sqrt{\mathbb{E}_{(X,Y)\sim\gamma}|\kappa(X,z)-\kappa(Y,z)|^2}$$

I.i.d. samples $N = m = n$, $\hat{\mu} = \sum_{i=1}^N \frac{1}{N}\delta_{x_i}$ and $\hat{\nu} = \sum_{i=1}^N \frac{1}{N}\delta_{y_i}$

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

$l = 2N$ evaluations each requires $\mathcal{O}(N\log N)$

```
% 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);
witness_func = @(x) kappa(x, Z{i_star})
```

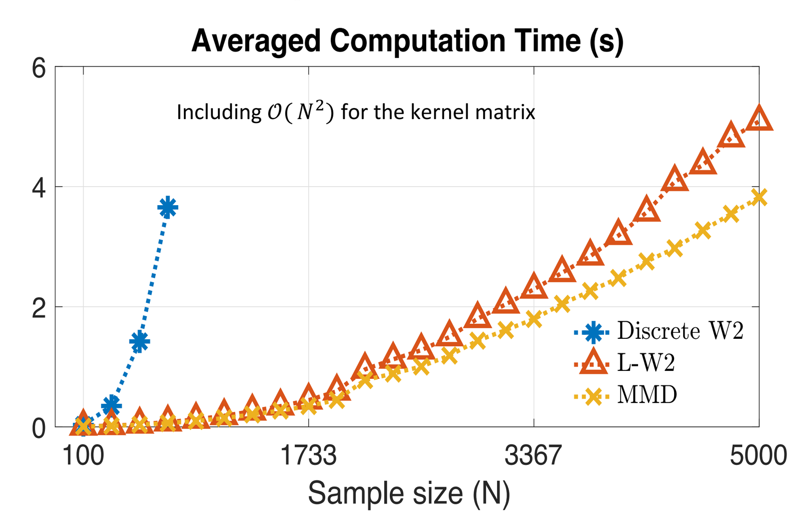
$$\omega(\cdot) = \kappa(\cdot, z_{i^\star})$$ Witness function

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

At most $l = 2N$ evaluations each requires $\mathcal{O}(N)$

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

Scalability tests: L-W2, MMD, discrete W2

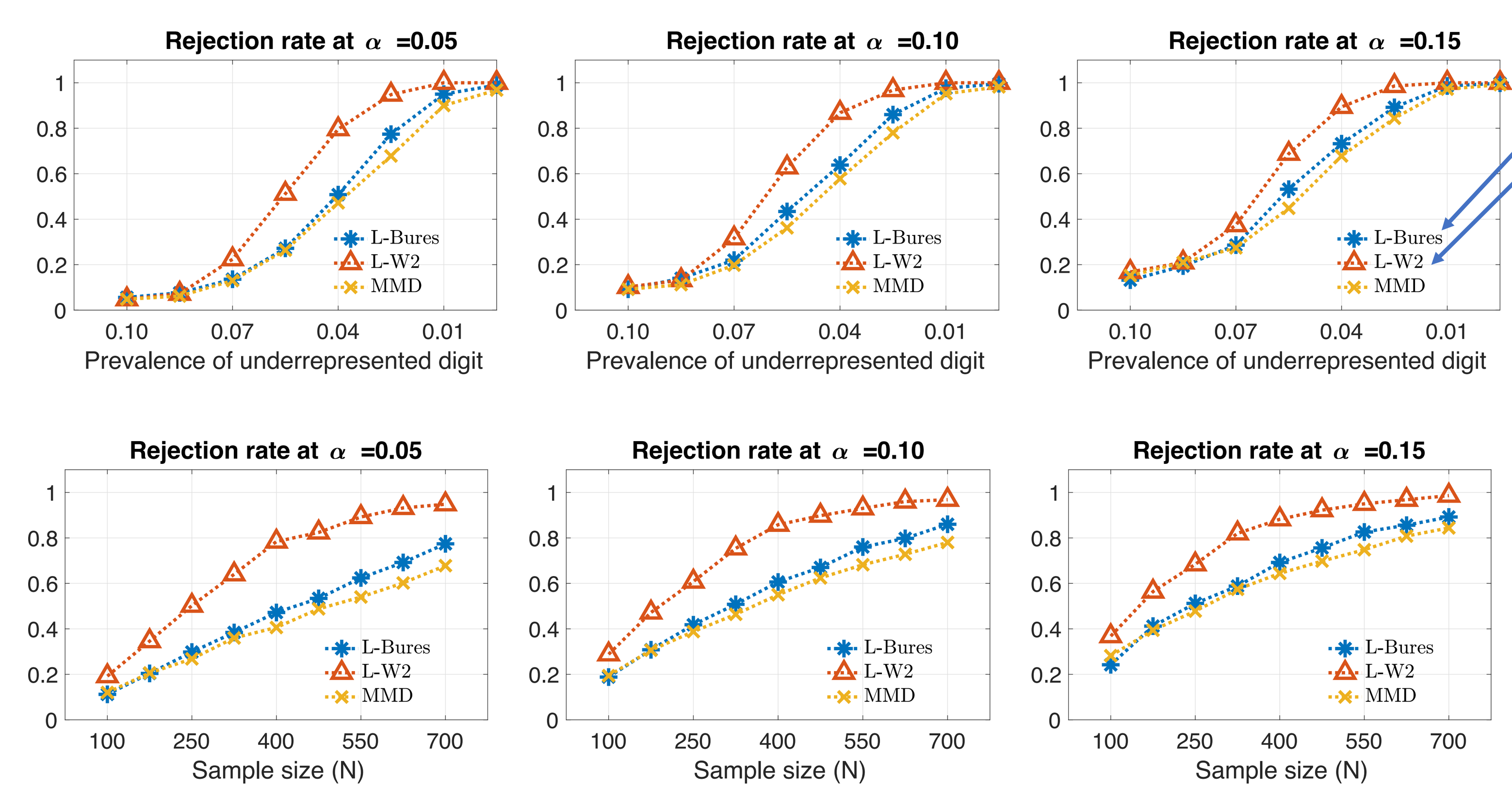


We compare computation time of our method, MMD, and Wasserstein. Computation time is averaged over 10 digits. The complexity of discrete Wasserstein distance is $\mathcal{O}(N^3)$ whereas our proposed method is only $\mathcal{O}(N^2\log(\mathrm{N}))$.

Covariate Shift Detection

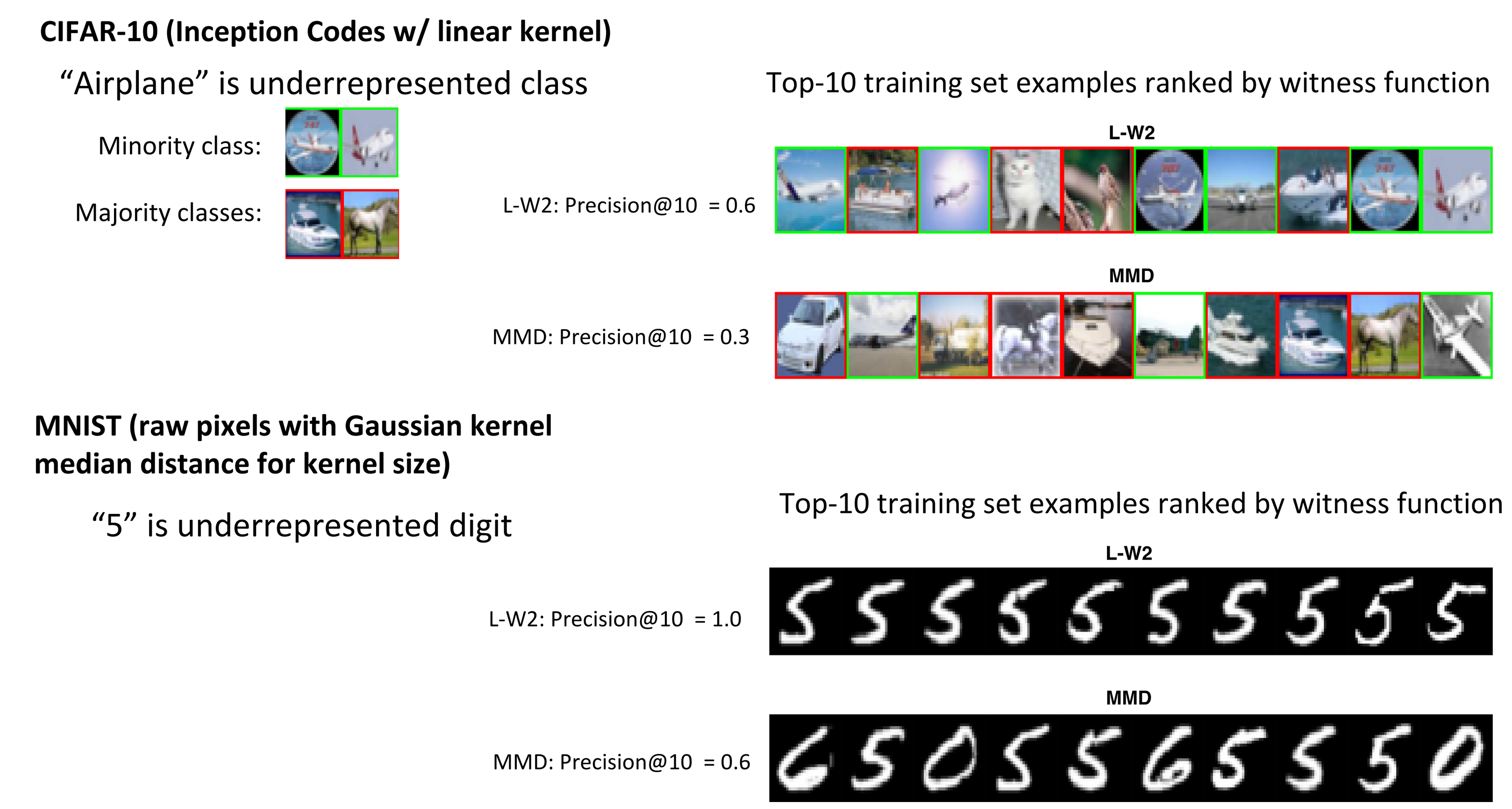
We perform a statistical power test to detect the difference between a sample with a uniform distribution of classes and a sample with the underrepresented class.

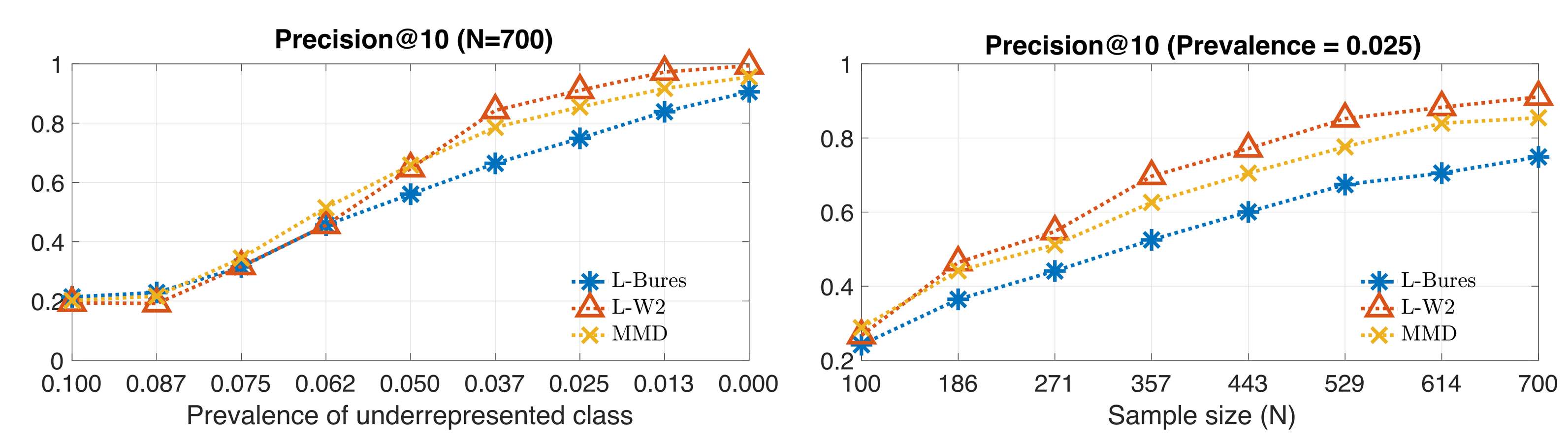
Statistical power test as a function of the class prevalence and sample size on MNIST dataset for digit "0"



Identify the Class Imbalance with Witness Function

Precision of the witness function in detecting underrepresented classes

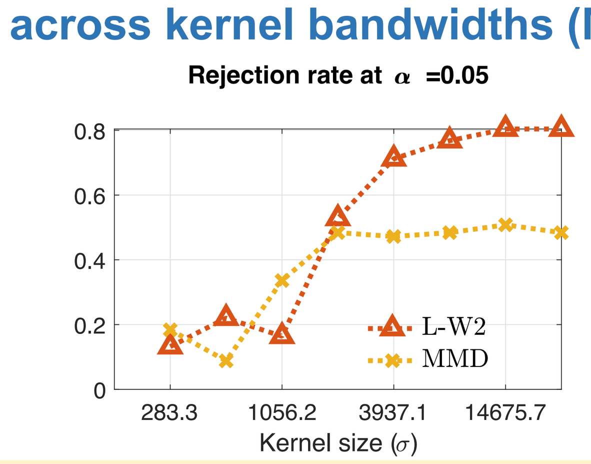




Averaged-precision@10 on MNIST dataset where the minority class is "6". The precision@10 was calculated by averaging 500 Monte Carlo samples iterations. Landmark-based kernel Bures (L-Bures), landmark kernel-Wasserstein (L-W2) and MMD divergences. (Left) The sample size is 700 for each set. (Right) The prevalence of the underrepresented digit is 0.025.

L-Bures: Landmark max-sliced kernel Bures
L-W2: Landmark max-sliced kernel Wasserstein

Power test across kernel bandwidths (MNIST digit "4")



The statistical power as a function of the kernel bandwidth. We obtained a priori global "median" bandwidth. Then we applied the power test on range of kernels sizes in which the priori bandwidth is centered. Sample size is 500 and the underrepresented class's prevalence is 0.025. Instances in each sample are randomly permuted between the two samples for 150 times with 250 Monte Carlo samples iterations.

The implementation of our approach and demos can be found by scanning QR-code