Kernel Landmarks: An Empirical Statistical Approach to Detect Covariate Shift

Yüksel Karahan
University of Delaware
Newark, Delaware, USA
ykarahan@udel.edu

Bilal Riaz
University of Delaware
Newark, Delaware, USA
bilalria@udel.edu

Austin J. Brockmeier[*]
University of Delaware
Newark, Delaware, USA
ajbrock@udel.edu

Abstract

Training an effective predictive model with empirical risk minimization requires a distribution of the input training data that matches the testing data. Covariate shift can occur when the testing cases are not class-balanced, but the training is. In order to detect when class imbalance is present in a test sample (without labels), we propose to use statistical divergence based on the Wasserstein distance and optimal transport. Recently, slicing techniques have been proposed that provide computational and statistical advantages for the Wasserstein distance for high-dimensional spaces. In this work we presented a computationally simple approach to perform generalized slicing of the kernel-based Wasserstein distance and apply it as a two-sample test. The proposed landmark-based slicing chooses a single point to be the sole support vector to represent the witness function. We run pseudo-real experiments using the MNIST dataset and compare our method with maximum mean discrepancy (MMD). We have shown that our proposed methods perform better than MMD on these synthetic simulations of covariate shift.

1 Introduction

Optimal transport (OT) is an old problem that can be dated back to 17th century. Historically the object of interest was literal dirt, but today we can apply the same approach to align two feature distributions or data sets [Peyré and Cuturi, 2019]. Generally, computing Wasserstein distance requires solving a high-dimensional optimization problem. There have been many extensions and approximations of optimal transport. Perhaps the two easiest cases to compute Wasserstein distance are for one-dimensional distributions or distributions described by their first and second-moments, as in Gaussian distributions. However, most data sets have multivariate features and are non-Gaussian. Using the Radon transform, Kolouri et al. [2016] proposed to obtain slices that are one-dimensional marginal representations of the distributions through linear projections. The max-sliced Wasserstein [Deshpande et al., 2019, Nguyen et al., 2021] replaces the random projections with a single slice or an optimized distribution over slices that maximizes the 1D Wasserstein. Alternatively, Meng et al. [2019] proposed to find the most meaningful projection using a sufficient dimension reduction technique [Li, 1991].

We propose an alternative solution called *kernel landmarks* to tackle this problem. We use a kernel-based approach, mapping the data points to a Hilbert space [Zhang et al., 2020] and evaluate the discrepancy between the distributions using each data point as a witness function and measure the divergence between the witness function evaluations. After evaluating all data points, we pick the point (the landmark) which identifies the largest discrepancy between the distribution. Our preliminary results show that this landmark-based max slicing is nearly efficient as maximum mean discrepancy (MMD) [Gretton et al., 2012]. Furthermore, the landmark-based kernel max-slicing

[*] `https://www.eecis.udel.edu/~ajbrock`

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

is much simpler to compute than kernel max-slicing [Brockmeier et al., 2021], while still being a probability distance metric.

2 Methodology

We consider a feature domain $\mathcal{X} \subseteq \mathbb{R}^d$. Let $P(\mathcal{X})$ be the set of Borel probability measures on the metric space $(\mathcal{X}, d)$ where $d(x, y)$ is the distance metric for $x, y \in \mathcal{X}$. Let $\mu, \nu \in P(\mathcal{X})$ be the probability measures and $X, Y \in \mathcal{X}$ be the random variables such that $X \sim \mu$ and $Y \sim \nu$. For any $p \geq 1$, we assume the distributions μ and ν have finite p-th moments, and using the Euclidean distance, $d(x, y) = \|x - y\|$, the Wasserstein-p distance is given as

$$W_p(\mu, \nu) = \Big[\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \Big]^{\frac{1}{p}}. \tag{1}$$

Cédric [2003] shows that Eq. (1) gives a metric on $P(\mathcal{X})$ [Villani, 2008, Peyré and Cuturi, 2019]. The max-sliced Wasserstein-p distance is given by the following saddlepoint problem

$$W_p^*(\mu, \nu) = \sup_{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 = 1} \left[\inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma} |\langle X - Y, \mathbf{w} \rangle|^p \right]^{\frac{1}{p}}, \tag{2}$$

where $\|(\mathbf{w}\mathbf{w}^\top)(X - Y)\| = \|\langle X - Y, \mathbf{w}\rangle \mathbf{w}\| = |\langle X - Y, \mathbf{w}\rangle| \, \|\mathbf{w}\|$.

2.1 Divergences in the Reproducing Kernel Hilbert Space (RKHS)

Consider a symmetric (real-valued) positive definite kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. κ defines $\mathcal{H}$, a reproducing kernel Hilbert space, if the following conditions are met: $\forall x \in \mathcal{X},\ \kappa(\cdot, x) \in \mathcal{H}$ and $\forall x \in \mathcal{X},\ f \in \mathcal{H},\quad f(x) = \langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}}$ [Scholkopf and Smola, 2001]. Let $\phi : \mathcal{X} \to \mathcal{H}$ be the implicit feature map (mapping elements $x \in \mathcal{X}$ to elements in RKHS $\phi(x) = \kappa(\cdot, x)$) such that $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = \kappa(x, y)$.

Assuming a bounded kernel $\mathbb{E}_{X\sim\xi}[\kappa(X, X)] \leq \infty, \quad \forall \xi \in P(\mathcal{X})$, for bounded family of functions in the RKHS $\mathcal{F} = \{\omega : \langle \omega, \omega\rangle_{\mathcal{H}} \leq 1\}$ on $\mathcal{X}$ where $\omega : \mathcal{X} \to \mathbb{R}$, the maximum mean discrepancy (MMD) is given by

$$\mathrm{MMD}^{\mathcal{H}}(\mu, \nu) = \sup_{\omega \in \mathcal{F}} \mathbb{E}_{X\sim\mu, Y\sim\nu}[\langle \phi(X) - \phi(Y), \omega\rangle] = \sup_{\omega \in \mathcal{F}} \mathbb{E}[\omega(X) - \omega(Y)] = \|m_\mu - m_\nu\|_{\mathcal{H}}, \tag{3}$$

where $m_\mu = \mathbb{E}_{X\sim\mu}[\phi(X)] \in \mathcal{H}$ and $m_\nu = \mathbb{E}_{Y\sim\mu}[\phi(Y)] \in \mathcal{H}$.

Using the kernel-induced distance $d(x, y) = \|\phi(x) - \phi(y)\|_{\mathcal{H}}$, Eq. (1) can be extended to the kernel Wasserstein-p distance [Zhang et al., 2020] as

$$W_p^{\mathcal{H}}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\gamma} \left[\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^p \right] \right)^{\frac{1}{p}}, \tag{4}$$

where $\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^p = (\kappa(X, X) - 2\kappa(X, Y) + \kappa(Y, Y))^{p/2}$. For $p = 2$, this simplifies such that the joint expectation moves inside and $W_2^{\mathcal{H}}(\mu, \nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \sqrt{\mathbb{E}[\kappa(X, X)] + \mathbb{E}[\kappa(Y, Y)] - \mathbb{E}_{(X,Y)\sim\gamma} 2\kappa(X, Y)}$. If one were to pursue the one-dimensional case as the easiest way to compute Wasserstein distance, the implicit feature map $\phi(\cdot)$ seems to be an obstacle to slicing compare to Eq. (2). However, by the reproducing property of the RKHS, the sliced distance defined by $\omega \in \mathcal{H}$, $\|\omega\|_{\mathcal{H}} = 1$, is $\|(\omega \otimes \omega)(\phi(X) - \phi(Y))\|_{\mathcal{H}} = |\langle \phi(X) - \phi(Y), \omega\rangle| \|\omega\|_{\mathcal{H}} = |\omega(X) - \omega(Y)|$, where $\omega(X)$ and $\omega(Y)$ are real-valued random variables with pushforward measures $\omega_\sharp\mu$ and $\omega_\sharp\nu$, respectively. The max-sliced kernel Wasserstein-2 distance can be expressed in terms of witness functions as

$$W_2^{\mathcal{H}_*}(\mu, \nu) = \sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} = 1} W_2(\omega_\sharp\mu, \omega_\sharp\nu) = \sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} = 1} \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\mathbb{E}_{(X,Y)\sim\gamma} \left[|\omega(X) - \omega(Y)|^2 \right] \right)^{\frac{1}{2}} \tag{5}$$

$$= \sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} = 1} \left(\mathop{\mathbb{E}}_{X\sim\mu}[\omega^2(X)] + \mathop{\mathbb{E}}_{Y\sim\nu}[\omega^2(Y)] - \sup_{\gamma \in \Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{(X,Y\sim\gamma)}[2\omega(X)\omega(Y)] \right)^{\frac{1}{2}}. \tag{6}$$

From the second line it is clear that the optimal γ can be solved analytically by coupling the largest values of $\omega(X)$ with the largest values of $\omega(Y)$, this can be done using their inverse cumulative distribution functions, which exist under mild conditions [Santambrogio, 2015]. Even with analytic solutions for the optimal transport after slicing, the computation of the max-sliced Wasserstein distance still requires optimizing the parameters defining the optimal slice.

Here, we propose a restricted form of kernel max-slicing called kernel landmarks. We restrict the witness function to be an implicit mapping of a single data point in the Hilbert space, $\omega = \phi(z), z \in \mathcal{X}$, assuming normalized kernel $\kappa(z, z) = 1, \quad z \in \mathcal{X} \implies \|\phi(z)\|_{\mathcal{H}} = 1$, and compute the Wasserstein-2 distance

$$W_2^{\mathcal{H}_{L^*}}(\mu, \nu) = \sup_{z \in \mathcal{X}} \inf_{\gamma \in \Gamma(\mu,\nu)} \sqrt{\mathbb{E}_{(X,Y)\sim\gamma} |\kappa(X, z) - \kappa(Y, z)|^2}. \tag{7}$$

For characteristic kernel functions [Fukumizu et al., 2008], namely the Gaussian and Laplacian kernels, this is a probability metric as stated and proved in the appendix (Theorem 1).

2.2 Two-sample Tests Using Kernel Divergences

We now consider two finite, weighted samples $\{(\mu_i, x_i)\}_{i=1}^m$ and $\{(\nu_i, y_i)\}_{i=1}^n$ of size m and n with $\sum_i \mu_i = 1$ and $\sum_i \nu_i = 1$. The masses are represented by the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. These samples can be represented by the empirical measures $\hat{\mu} = \sum_i \mu_i \delta_{x_i}$ and $\hat{\nu} = \sum_i \nu_i \delta_{y_i}$.

In the sample case, the witness function ω is parametrized as $\omega(\cdot) = \sum_{i=1}^l \alpha_i \kappa(\cdot, z_i)$ in terms of the dual variables $\boldsymbol{\alpha} \in \mathbb{R}^l$ and $\{z_i\}_{i=1}^l = \mathcal{Z}$ where $z_i = \begin{cases} x_i, & 1 \leq i \leq m \\ y_{i-m}, & m+1 \leq i \leq l \end{cases}, i \in \{1, \ldots, l\}$ and $l = m + n$. The kernel matrix is $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{XZ} \\ \mathbf{K}_{YZ} \end{bmatrix} \in \mathbb{R}^{l \times l}$ where $\mathbf{K}_{XZ} \in \mathbb{R}^{m \times l}$,$\mathbf{K}_{YZ} \in \mathbb{R}^{n \times l}$, and $K_{i,j} = \kappa(z_i, z_i) - 2\kappa(z_i, z_j) + \kappa(z_j, z_j)$ for $i \in \{1, \ldots, m+n\}$.

The witness function evaluations for each sample are given by the vectors $[\omega(x_1), \ldots, \omega(x_m)]^\top = \mathbf{K}_{XZ}\boldsymbol{\alpha}$ and $[\omega(y_1), \ldots, \omega(y_n)]^\top = \mathbf{K}_{YZ}\boldsymbol{\alpha}$. For a positive definite kernel matrix $\mathbf{K}$, to ensure ω has bounded norm, the coefficient vector should be restricted to be $\boldsymbol{\alpha} \in \mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^l : \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \leq 1\}$. Additional constraints are needed if the kernel matrix is positive semi-definite.

The max-sliced kernel Wasserstein-2 distance (squared) can be computed in terms of witness function evaluations as

$$\begin{aligned} W_2^{\mathcal{H}_*}(\hat{\mu}, \hat{\nu})^2 =& \max_{\boldsymbol{\alpha} \in \mathcal{A}} \min_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \left\{ \sum_{i,j} P_{ij} |\omega(x_i) - \omega(y_j)|^2 = \langle \mathbf{P}, (\mathbf{K}_{XZ}\boldsymbol{\alpha}\mathbf{1}_n^\top - \mathbf{1}_m(\mathbf{K}_{YZ}\boldsymbol{\alpha})^\top)^{\circ 2} \rangle \right\} \\ =& \max_{\boldsymbol{\alpha} \in \mathcal{A}} \langle \boldsymbol{\mu}, (\mathbf{K}_{XZ}\boldsymbol{\alpha})^{\circ 2} \rangle + \langle \boldsymbol{\nu}, (\mathbf{K}_{YZ}\boldsymbol{\alpha})^{\circ 2} \rangle - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}\mathbf{K}_{YZ}\boldsymbol{\alpha}, \mathbf{K}_{XZ}\boldsymbol{\alpha} \rangle, \end{aligned} \tag{8}$$

where $\mathcal{P}_{\hat{\mu},\hat{\nu}} = \{\mathbf{P} \in [0, 1]^{m \times n} : \mathbf{P}\mathbf{1}_n = \boldsymbol{\mu} \text{ and } \mathbf{P}^\top \mathbf{1}_m = \boldsymbol{\nu}\}$ is the transport polytope, where $\mathbf{1}_n^\top = [1, \ldots, 1]$ is a vector of n ones, and $\mathbf{M}^{\circ 2}$ is the element-wise squaring of the entries of $\mathbf{M}$. The optimal transport plan can be obtained analytically after sorting the values in the vectors $\boldsymbol{\omega}_Y = \mathbf{K}_{YZ}\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{\omega}_X = \mathbf{K}_{YZ}\boldsymbol{\alpha} \in \mathbb{R}^n$. Assuming Q and R are the permutations such that $\omega(x_{Q(1)}) \leq \cdots \leq \omega(x_{Q(m)})$ and $\omega(y_{R(1)}) \leq \cdots \leq \omega(y_{R(n)})$. The optimal transport plan $\mathbf{P}_\star$ is defined as $[\mathbf{P}_\star]_{i,j} = [\tilde{\mathbf{P}}]_{Q(i),R(j)}$ where $\tilde{\mathbf{P}}$ is optimal transport plan after sorting, which is the finite differences across rows and columns of $\mathbf{G}$: $[\tilde{\mathbf{P}}]_{i,j} = G_{i+1,j+1} - G_{i+1,j} - G_{i,j+1}$ with $G_{i,j} = \min(\sum_{k=1}^{i-1} \mu_{Q(k)}, \sum_{k=1}^{j-1} \nu_{R(k)}$, where the matrix $\mathbf{G} \in [0, 1]^{m+1 \times n+1}$ is the zero-padded joint cumulative distribution of the optimal transport plan after sorting. Overall, it is a saddle-point optimization problem, with evaluation cost $\mathcal{O}(N \log(N))$ for $n > m$ where $N = max(m, n)$.

For the kernel landmark Wasserstein distance (L-W2), the continuous multivariate optimization of the optimal slice is replaced by a discrete optimization over the set of possible landmarks is $\{z_i\}_{i=1}^l = \mathcal{Z}$. This is equivalent to restricting $\boldsymbol{\alpha}$ to be one-hot vector, e.g., $\boldsymbol{\alpha} = [0, \ldots, 0, 1, 0, \ldots, 0]^\top$. Assuming the i-th data point is the landmark, $\omega(\cdot) = \kappa(\cdot, z_i)$, $\mathbf{K}_{XZ}\boldsymbol{\alpha} = [\kappa(x_1, z_i), \ldots, \kappa(x_m, z_i)]^\top = \mathbf{k}_{Xz_i}$, and $\mathbf{K}_{YZ}\boldsymbol{\alpha} = [\kappa(y_1, z_i), \ldots, \kappa(y_n, z_i)]^\top = \mathbf{k}_{Yz_i}$, which are the i-th columns of $\mathbf{K}_{XZ}$ and $\mathbf{K}_{YZ}$, respectively. Using these, we introduce the landmark-based max-sliced kernel Wasserstein-2 distance

$$\begin{aligned} W_2^{\mathcal{H}_{L^*}}(\hat{\mu},\hat{\nu}) &= \max_{k\in\{1,\ldots,l\}} \min_{\mathbf{P}\in\mathcal{P}_{\hat{\mu},\hat{\nu}}} \sqrt{\sum_{i,j} P_{ij}|\kappa(x_i,z_k)-\kappa(y_j,z_k)|^2} \\ &= \max_{i\in\{1,\ldots,l\}} \sqrt{\langle\boldsymbol{\mu},\mathbf{k}_{Xz_i}^{\circ 2}\rangle + \langle\boldsymbol{\nu},\mathbf{k}_{Yz_i}^{\circ 2}\rangle - 2\max_{\mathbf{P}\in\mathcal{P}_{\hat{\mu},\hat{\nu}}}\langle\mathbf{P}\mathbf{k}_{Yz_i},\mathbf{k}_{XZ_i}\rangle}. \end{aligned} \tag{9}$$

We now assume i.i.d. samples, $\mu_1 = \cdots = \mu_m = \frac{1}{m}$ and $\nu_1 = \cdots = \nu_m = \frac{1}{n}$. Let $\mathbf{P}_i$ be the optimal transport matrix for landmark $i \in \{1,\ldots,l\}$. It can be written as row and column permutations of $\tilde{\mathbf{P}}$ the optimal transport matrix for sorted i.i.d. samples: $\mathbf{P}_i = \mathbf{Q}_i^\top\tilde{\mathbf{P}}\mathbf{R}_i$, $\mathbf{Q}_i \in \Pi_m, \mathbf{R}_i \in \Pi_n$, where the entries of $\mathbf{Q}_i\mathbf{k}_{Xz_i} = [\kappa(x_{Q_i(1)},z_i),\ldots,\kappa(x_{Q_i(m)},z_i)]^\top$ and $\mathbf{R}_i\mathbf{k}_{Yz_i} = [\kappa(y_{R_i(1)},z_i),\ldots,\kappa(y_{R_i(n)},z_i)]$ are in ascending order, $\kappa(x_{Q_i(1)},z_i) \le \cdots \le \kappa(x_{Q_i(m)},z_i)$ and $\kappa(y_{R_i(1)},z_i) \le \cdots \le \kappa(y_{R_i(n)},z_i)$. ($\mathbf{Q}_i$ and $\mathbf{R}_i$ are the permutation matrices corresponding to permutations $Q_i(\cdot)$ and $R_i(\cdot)$, where Π_n denotes the set of $n\times n$ permutations matrices.) The sorted witness function evaluations for each landmark can be expressed together as $\tilde{\mathbf{K}}_{XZ} = [\mathbf{Q}_1\mathbf{k}_{Xz_1},\ldots,\mathbf{Q}_l\mathbf{k}_{Xz_l}] = [\tilde{\mathbf{k}}_{Xz_1},\ldots,\tilde{\mathbf{k}}_{Xz_l}]$ and $\tilde{\mathbf{K}}_{YZ} = [\mathbf{R}_1\mathbf{k}_{Yz_1},\ldots,\mathbf{R}_l\mathbf{k}_{Yz_l}] = [\tilde{\mathbf{k}}_{Yz_1},\ldots,\tilde{\mathbf{k}}_{Yz_l}]$.

$$W_2^{\mathcal{H}_{L^*}}(\hat{\mu},\hat{\nu}) = \sqrt{\max_{i\in\{1,\ldots,l\}}\left[\frac{1}{m}\|\mathbf{K}_{XZ}\|_2^2 + \frac{1}{n}\|\mathbf{K}_{YZ}\|_2^2 - 2\mathbf{1}_n^\top((\tilde{\mathbf{P}}^\top\tilde{\mathbf{K}}_{XZ})\circ\tilde{\mathbf{K}}_{YZ})\right]_i}. \tag{10}$$

In the case of equal sample sizes $m = n$ and $\mu_1 = \cdots = \mu_m = \nu_1 = \cdots = \nu_m = \frac{1}{m}$, the optimal transport matrix is a scaled product of permutation matrices $\mathbf{P}_i = \frac{1}{m}\mathbf{Q}_i^\top\mathbf{R}_i$, since $\tilde{\mathbf{P}} = \frac{1}{m}\mathbf{I}$. In this case, the kernel landmark Wasserstein-2 distance is simply

$$\max_{i\in\{1,\ldots,l\}}\frac{1}{\sqrt{m}}\|\tilde{\mathbf{k}}_{Xz_i} - \tilde{\mathbf{k}}_{Yz_i}\|_2 = \sqrt{\frac{1}{m}\max_{i\in\{1,\ldots,l\}}[\mathbf{1}_m^\top(\tilde{\mathbf{K}}_{XZ}-\tilde{\mathbf{K}}_{YZ})^{\circ 2}]_i}, \tag{11}$$

where the second expression computes the Euclidean norm of each column of the differences.

3 Covariate Shift Detection

In this section, we perform simulation experiments to compare kernel landmark Wasserstein-2 distance and MMD for detecting imbalanced classes, as a specific form of covariate shift detection. (We also compare a kernel landmark based Bures distance described in the Appendix.) 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. We also examine the ability of the witness function to identify instances associated with the underrepresented class.

Specifically, we consider the MNIST dataset split into train and test sets across different levels of imbalance and sample size. $\hat{\mu}$ represents a sample of the training set with a balanced proportion of each class, and $\hat{\nu}$ has less instances from one class. MNIST has ten class labels $L \in \{0,1,\ldots,9\}$. Let P_L be the prevalence of L-th class/digit, which is underrepresented in $\hat{\nu}$. $P_L = \frac{1}{10}(1-p)$ where $p\in[0,1]$ is the Bernoulli probability that the underrepresented digit is replaced by a majority class digit when it is drawn. The prevalence of the any other digit is expected to be $P_{L'\in\{0,1,\cdots,9\}\setminus L} = \frac{1}{10}(1-p)+\frac{1}{9}p$. For example, when p takes values of 0, 0.5, or 0.8, the probabilities of underrepresented digit are $\frac{1}{10}$, $\frac{1}{20}$, and $\frac{1}{50}$, respectively.

We use the kernel-based divergences to test the hypothesis that the two samples come from the same distribution. We use the Gaussian kernel $\kappa(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ where σ is median of the pairwise distances. As a significant threshold for the divergence we use the $1-\alpha$ quantile of the surrogate distribution of divergence values when instances in each sample are randomly permuted between the two samples (250 times). To estimate the statistical power for a given occurrence level, we use 500 Monte Carlo samples iterations. Fig. 1 and Fig. 3 show the statistical power for $L = 0$, digit "0", across different values of P_0 and sample size.

We also test the precision of the witness function in detecting the specific discrepancies associated to the class imbalance. We test whether the instances in the training set with the largest magnitude

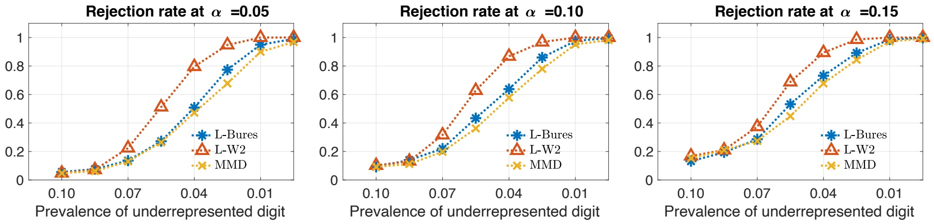

Figure 1: The figure illustrates the power test on MNIST dataset. We compare the Landmark-based kernel Bures,Wasserstein and MMD divergences. The power curves for three different critical values, α, as a function of prevalence of the underrepresented digit. For this experiment, the sample sizes are $m = n = 700$ and the digit is "0".

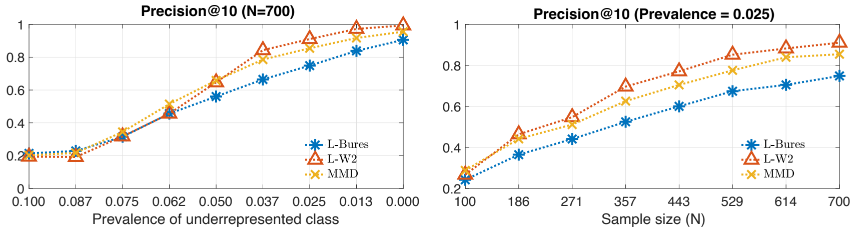

Figure 2: Averaged-precision@10 on MNIST dataset where the minority class is “6”. The precision@10 was calculated by averaging 500 Monte Carlo samples iterations. Landmark-based kernel Bures (L-Bures), landmark kernel-Wasserstein (L-W2) and MMD divergences. (Left) The sample size is $N = 700$ for each test and train set. (Right) The prevalence of the underrepresented digit is 0.025.

witness function evaluations are from the minority class. We report the evaluation results using precision at 10 in Fig. 2 for the specific case where digit “6” class in the test set. We also compare the proposed approach to MMD’s witness function evaluations, which are slightly less precise. This approach is useful in practice, since by examining the label distribution of the top-K training set examples, the user can understand if there is any imbalances in the test set. Examples of the instances with highest witness function for both the MNIST and for CIFAR-10 are shown in the Appendix. The statistical power across different kernel bandwidths (testing the sensitivity to the median heuristic) is also shown in the Appendix. The implementation of our approach and demos can be found at `https://github.com/drpointcloud/landmark`.

4 Conclusion

In this paper we have investigated max-slicing for the kernel-based Wasserstein distance to detect class-based covariate shift. Our approach evaluates the discrepancy between distributions in terms of the similarity to a landmark point. Unlike the generalized max-sliced Wasserstein distance, the proposed distance can be computed exactly and efficiently for the case of two samples. Statistical power tests are employed to evaluate the performance of detecting class imbalances in testing data. We compared our approach with MMD, which is a well-known approach to find the discrepancy. The preliminary results shows that the proposed method detects simple cases of covariate shift better than MMD.

Acknowledgments and Disclosure of Funding

Yüksel Karahan and Austin J. Brockmeier’s effort were sponsored by the Department of the Navy, Office of Naval Research under ONR award number N00014-21-1-2300. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research. Bilal Riaz is supported by the US-Pakistan Knowledge Corridor Ph.D. program at the University of Delaware.

References

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL `http://dx.doi.org/10.1561/2200000073`.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5258–5267, 2016. doi: 10.1109/CVPR.2016.568.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander Schwing. Max-sliced Wasserstein distance and its use for GANs, 2019.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=QYjO70ACDK`.

Cheng Meng, Yuan Ke, Jingyi Zhang, Mengrui Zhang, Wenxuan Zhong, and Ping Ma. Large-scale optimal transport map estimation using projection pursuit. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8118–8129, 2019.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Z. Zhang, M. Wang, and A. Nehorai. Optimal transport in reproducing kernel Hilbert spaces: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7): 1741–1754, 2020. doi: 10.1109/TPAMI.2019.2903050.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Austin J. Brockmeier, Claudio Cesar Claros, Carlos H. Mendoza-Cardenas, Yüksel Karahan, Matthew S. Emigh, and Luis Gonzalo Sanchez Giraldo. Max-sliced Bures distance for interpreting discrepancies, 2021. URL `https://openreview.net/forum?id=D2Fp_qheYu`.

Villani Cédric. *Topics in optimal transportation*. Graduate studies in mathematics. American mathematical society, Providence, Rhode Island, 2003. ISBN 0-8218-3312-X.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer International Publishing, 2015. doi: 10.1007/978-3-319-20828-2. URL `https://doi.org/10.10072F978-3-319-20828-2`.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. ISSN 0723-0869. doi: https://doi.org/10.1016/j.exmath.2018.01.002.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

A Appendix

The appendix contains the statement and proof of the metric property of the kernel landmark Wasserstein-2 distance; a description of the kernel landmark Bures divergence; and additional results.

Theorem 1. *If κ is characteristic [Fukumizu et al., 2008] and normalized $\kappa(z,z) = 1 \quad \forall z \in \mathcal{X}$, then $W_p^{\mathcal{H}_{L^*}}(\mu,\nu)$ is a probability distance metric. For $\mu,\nu,\xi \in P(\mathcal{X})$ with finite p-th moments*

- $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) \geq 0$
- $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) = W_p^{\mathcal{H}_{L^*}}(\nu,\mu)$
- $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) = 0 \iff \mu = \nu.$
- $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) \leq W_p^{\mathcal{H}_{L^*}}(\mu,\xi) + W_p^{\mathcal{H}_{L^*}}(\nu,\xi).$

Proof. Non-negativity and symmetry are obvious from the proprieties of the Wasserstein distance. If $\mu = \nu$, then for any $\omega \in \mathcal{H}$, $\inf_{\gamma\in\Gamma(\mu,\mu)} \mathbb{E}_{(X,Y)\sim\gamma}|\omega(X) - \omega(Y)|^p = 0$.

We assume $\mu \neq \nu$ and proceed to lower bound the distance and show that $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) > 0$.

For any $\omega \in \mathcal{H}$,

$$\begin{aligned}\inf_{\gamma\in\Gamma(\mu,\mu)} \mathbb{E}_{(X,Y)\sim\gamma}|\omega(X) - \omega(Y)|^p &\geq \inf_{\gamma\in\Gamma(\mu,\mu)} |\mathbb{E}_{(X,Y)\sim\gamma}[\omega(X) - \omega(Y)]|^p && (12)\\ &= |\langle m_\mu - m_\nu, \omega\rangle|^p, && (13)\end{aligned}$$

where Jensen's inequality is used based on the convexity of $|\cdot|^p$. Taking the supremum over the set of landmarks yields an expression in terms of the difference of the means in the RKHS

$$\sup_{z\in\mathcal{X}} \inf_{\gamma\in\Gamma(\mu,\mu)} \mathbb{E}_{(X,Y)\sim\gamma}|\omega(X) - \omega(Y)|^p \geq \sup_{z\in\mathcal{X}} |\langle \underbrace{m_\mu - m_\nu}_{h\in\mathcal{H}}, \phi(z)\rangle|_p = \sup_{z\in\mathcal{X}} |h(z)|^p. \qquad (14)$$

When κ is characteristic, the mapping $m_\xi : \xi \mapsto \mathbb{E}_{X\sim\xi}[\phi(X)]$ is injective for $\xi \in P(\mathcal{X})$ [Fukumizu et al., 2008], and $\mu \neq \nu \implies m_\mu \neq m_\mu \implies \exists z, m_\mu(z) - m_\mu(z) = h(z) \neq 0$. Together this yields $W_p^{\mathcal{H}_{L^*}}(\mu,\nu) \geq \sup_{z\in\mathcal{X}}|h(z)|^p > 0$ for $\mu \neq \nu$.

The triangle inequality follows from the fact that the Wasserstein distance itself is a metric.

$$\begin{aligned}W_p^{\mathcal{H}_{L^*}}(\mu,\nu) &= \sup_{\omega\in\{\phi(z)\in\mathcal{H}:z\in\mathcal{X}\}} W_p^{\mathbb{R}}(\omega_\sharp\mu, \omega_\sharp\nu) && (15)\\ &\leq \sup_{\omega\in\{\phi(z)\in\mathcal{H}:z\in\mathcal{X}\}} \left(W_p^{\mathbb{R}}(\omega_\sharp\mu, \omega_\sharp\xi) + W_p^{\mathbb{R}}(\omega_\sharp\nu, \omega_\sharp\xi)\right) && (16)\\ &\leq \left(\sup_{\omega\in\{\phi(z)\in\mathcal{H}:z\in\mathcal{X}\}} W_p^{\mathbb{R}}(\omega_\sharp\mu, \omega_\sharp\xi)\right) + \left(\sup_{\omega\in\{\phi(z)\in\mathcal{H}:z\in\mathcal{X}\}} W_p^{\mathbb{R}}(\omega_\sharp\nu, \omega_\sharp\xi)\right) && (17)\\ &= W_p^{\mathcal{H}_{L^*}}(\mu,\xi) + W_p^{\mathcal{H}_{L^*}}(\nu,\xi) && (18)\end{aligned}$$

□

A.1 Kernel Landmark Bures Distance

When the μ and ν are Gaussian distributions, the Wasserstein-2 (W2) distance can be computed analytically in terms of the first and the second moments [Peyré and Cuturi, 2019]. The kernel Gaussian W2 distance [Zhang et al., 2020] is defined as $W_G^{\mathcal{H}}(\mu,\nu) = \sqrt{\|m_\mu - m_\nu\|^2 + d_B(\Sigma_\mu,\Sigma_\nu)^2}$, where $d_B(\Sigma_\mu,\Sigma_\nu) = \left(\operatorname{tr}(\Sigma_\mu) + \operatorname{tr}(\Sigma_\nu) - 2\operatorname{tr}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})\right)^{\frac{1}{2}}$ [Bhatia et al., 2019], $\Sigma_\mu = \rho_\mu - m_\mu \otimes m_\mu$ and $\Sigma_\nu = \rho_\nu - m_\nu \otimes m_\nu$ are the covariance matrices of X and Y in the RKHS, and $\rho_\mu = \mathbb{E}_{X\sim\mu}[\phi(X) \otimes \phi(X)] \in \mathcal{H}$ and $\rho_\nu = \mathbb{E}Y \sim \nu[\phi(Y) \otimes \phi(Y)] \in \mathcal{H}$ are the uncentered second moments. The Bures distance between the uncentered second moments in the Hilbert space [Brockmeier et al., 2021] is also a divergence measure $D_B^{\mathcal{H}}(\mu,\nu) = d_B(\rho_\mu,\rho_\nu)$.

For comparison, we consider a landmark-based version of the kernel-based max-sliced Bures distance. The kernel-based max-sliced Bures distance is expressed in terms of the squared witness functions as $D_B^{\mathcal{H}_*}(\mu,\nu) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1}\left|\sqrt{\mathbb{E}_{X\sim\mu}[\omega^2(X)]} - \sqrt{\mathbb{E}_{Y\sim\nu}[\omega^2(Y)]}\right|$. For i.i.d. samples, the kernel max-sliced Bures distance is $D_B^{\mathcal{H}_*}(\hat{\mu},\hat{\nu}) = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}\leq 1}\left|\frac{1}{\sqrt{m}}\|\mathbf{K}_{XZ}\boldsymbol{\alpha}\|_2 - \frac{1}{\sqrt{n}}\|\mathbf{K}_{YZ}\boldsymbol{\alpha}\|_2\right|$. The landmark-based max-sliced kernel Bures distance (L-Bures) is

$$D_B^{\mathcal{H}_{L^*}}(\hat{\mu},\hat{\nu}) = \max_{i\in\{1,\ldots,l\}}\left\{\left|\frac{1}{\sqrt{m}}\|\mathbf{k}_{Xz_i}\|_2 - \frac{1}{\sqrt{n}}\|\mathbf{k}_{Yz_i}\|_2\right|\right\}. \tag{19}$$

A.2 Additional Experiments

We performed statistical power test for various sample sizes (see Fig3) and kernel bandwidths (see Fig. 6). We also applied the proposed method and MMD on the purposefully imbalanced subsets of CIFAR10 dataset (please see Fig. 4) where each instance is represented by the internal representation of the inception network [Szegedy et al., 2016]: a 2048-dimensional vector. We used witness function evaluations to identify instances associated with the underrepresented class. In Fig. 7, we report the computation complexity of our method compared to MMD and the discrete Wasserstein-2 distance. As it can be seen in Fig. 7, the proposed method, is much easier to compute than the Wasserstein distance.

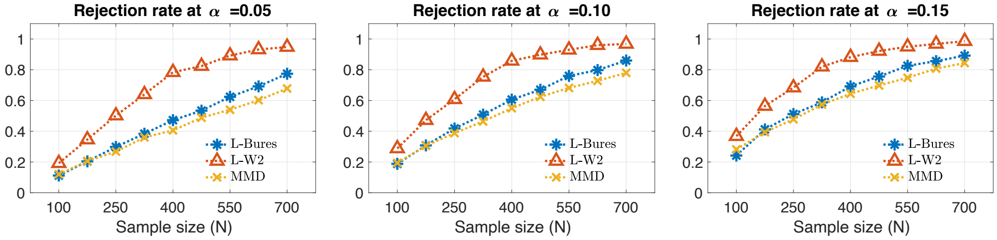

Figure 3: The figure illustrates the power test on MNIST dataset. We compare the Landmark-based kernel Bures, Wasserstein, and MMD divergences. The power curves for three different critical values, α, as a function of sample sizes. For this example, the prevalence of the underrepresented digit is 0.029 and the digit is "0". Instances in each sample are randomly permuted between the two samples for 250 times with 500 Monte Carlo samples iterations.

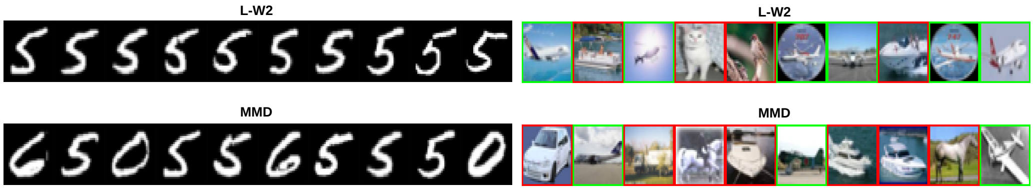

Figure 4: We used MNIST and CIFAR10 dataset to show the instances correspond to the top-10 largest values of witness function evaluations. We evaluate the performance of witness function to detect the instances from the minority class. (Left) The digit “5” is the minority class. Using the prevalence 0.025 and sample size 700, the top-10 witness function evaluations of our method identifies missing class instances with a high precision. (Right) We also compare the proposed approach and MMD to detect mismatched distributions of test and train images on the CIFAR-10 data set using the internal representation of the Inception Network. An instance is represented by a size of 2048 vector. In this case the minority class is “airplane”, the prevalence of missing class is 0.025, and the sample size is 700 for each set.

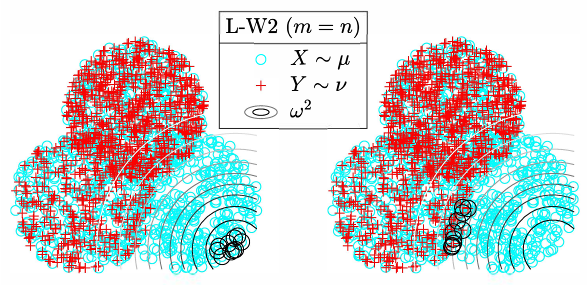

Figure 5: In this figure, the largest magnitude of the witness function evaluations is shown for Landmark max-sliced kernel Wasserstein (L-W2) which depict the discrepancies between distributions.

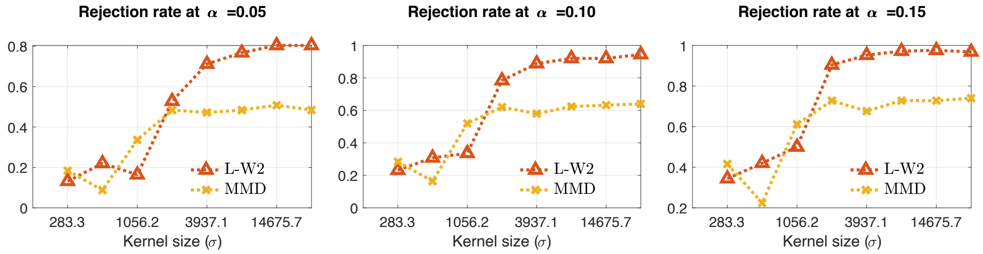

Figure 6: The figure illustrates the power test on MNIST dataset. We compare the Landmark Wasserstein and MMD divergences. The power curves for three different critical values, α, as a function of kernel bandwidths. For this example, the prevalence of the underrepresented digit ("4") is 0.025. Instances in each sample are randomly permuted between the two samples for 150 times with 250 Monte Carlo samples iterations.

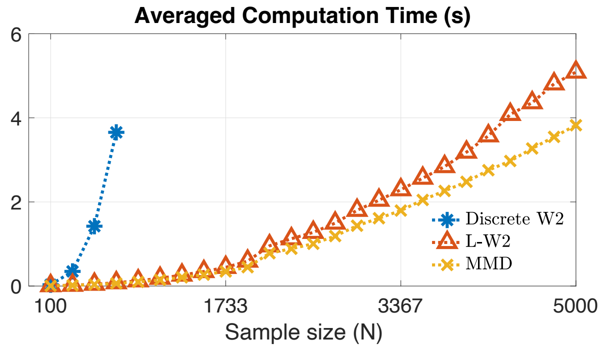

Figure 7: We compare computation time of our method, MMD, and Wasserstein Distance using MNIST dataset. Computation time is averaged over 10 digits. The complexity of Wasserstein distance is $\mathcal{O}(N^3)$ whereas our proposed method is only $\mathcal{O}(N^2 \log(N))$. As it can be seen our method which is an approximation of Wasserstein distance is much faster than Wasserstein distance. We did not add kernel max-sliced Wasserstein distance here because we can not directly compare its run time since its dual variable is obtained iteratively.