Kernel Landmarks: An Empirical Statistical Approach to Detect Covariate Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Training a 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 in high-dimensional spaces. In this work we presented a computationally simple approach to perform generalized slicing via kernel-based Wasserstein distance and apply it to as a two-sample test. The proposed landmark-based slicing chooses a single point in the samples 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 distributions or samples of points [9]. Generally, computing Wasserstein distance requires solving an high-dimensional optimization problem. There have been many extensions and approximations of optimal transport. Perhaps the two easiest cases to compute Wasserstein distance are when the data is one-dimensional or can be described by the first and second-moments, as in Gaussian distributions. However, data-sets usually have higher dimensions and are non-Gaussian. Inspired by the Radon transform, [6] proposed to obtain slices that are one-dimensional marginal representations of the distributions through linear projections. The max-sliced Wasserstein [4] was proposed to solve projection complexity by replacing the random projections with the one that maximized the 1D Wasserstein. Computing max-sliced Wasserstein still requires optimization and it could be expensive. Alternatively, [8] proposed to find the most meaningful projection using a sufficient dimension reduction technique [7].

We propose an alternative solution called *landmark-based kernel max-slicing* to tackle this problem. We use a kernel-based approach, mapping the data points to a Hilbert space [12] 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 efficient compared to maximum mean discrepancy (MMD) [5]. Furthermore, the landmark-based kernel max-slicing is much simpler to compute than kernel max-slicing [2].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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 a given metric space $(\mathcal{X}, d)$ 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 $p \geq 1$, the Wasserstein-p distance is given as

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

Villani [3] shows that square root of the objective in Eq.1 gives a metric on $P(\mathcal{X})$ and this metric is known as Wasserstein-2 distance [11, 9].

2.1 Divergences in the Reproducing Kernel Hilbert Space (RKHS)

Let $\mathcal{H}$ be the Hilbert space with the real-valued function and κ be a positive definite kernel such that $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is a reproducing kernel Hilbert space if the following conditions are met: $(i) \forall x \in \mathcal{X},\ \kappa(\cdot, x) \in \mathcal{H}$ and $(ii) \forall x \in \mathcal{X},\ f \in \mathcal{H}\ \langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ [10]. Let $\phi : \mathcal{X} \to \mathcal{H}$ be the feature map (mapping elements $x \in \mathcal{X}$ to elements in RKHS $\phi(x)$) such that $\phi(x) = \kappa(\cdot, x)$ and $\phi(y) = \kappa(\cdot, y)$, and $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \kappa(x, y)$.

For a family of functions $\mathcal{F} = \{\omega : \langle \omega, \omega \rangle_{\mathcal{H}} \leq 1\}$ on $\mathcal{X}$ and $\omega : \mathcal{X} \to \mathbb{R}$, Maximum Mean Discrepancy (MMD) is given by

$$\text{MMD}^{\mathcal{H}}(\mu, \nu) = \sup_{\omega \in \mathcal{F}} \mathbb{E}[\langle \phi(X) - \phi(Y), \omega \rangle] = \|m_\mu - m_\nu\|_{\mathcal{H}}. \tag{2}$$

$m_\mu = \mathbb{E}_{X \sim \mu}[\phi(X)] \in \mathcal{H}$ and $m_\nu = \mathbb{E}_{Y \sim \mu}[\phi(Y)] \in \mathcal{H}$ are the first moments of data on RKHS. MMD is the optimal transport in the RKHS between distributions described by their means.

When the μ and ν are two Gaussians, the Wasserstein distance can be related to MMD and Bures distance [9] in terms of the first and the second moments [12]. The kernel Gaussian W2 distance is defined as $W_G^{\mathcal{H}}(\mu, \nu) = \sqrt{\|m_\mu - m_\nu\|^2 + d_B^{\mathcal{H}}(\Sigma_\mu, \Sigma_\nu)^2}$, where $d_B(\Sigma_\mu, \Sigma_\nu) = \big(\text{Tr}(\Sigma_\mu) + \text{Tr}(\Sigma_\nu) - 2\text{Tr}(\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}})\big)^{\frac{1}{2}}$ [1], $\Sigma_\mu = \rho_\mu - m_\mu \otimes m_\mu$ and $\Sigma_\nu = \rho_\nu - m_\nu \otimes m_\nu$ are the covariance matrices of X and Y in the RKHS, and $\rho_\mu = \mathbb{E}[\phi(X) \otimes \phi(X)] \in \mathcal{H}$ and $\rho_\nu = \mathbb{E}[\phi(Y) \otimes \phi(Y)] \in \mathcal{H}$ are the uncentered second moments. The Bures distance between the uncentered second moments in the Hilbert space [2] is also a divergence measure $D_B^{\mathcal{H}}(\mu, \nu) = d_B(\rho_\mu, \rho_\nu)$.

The kernel Wasserstein-2 distance is

$$\inf_{\gamma \in \Gamma(\mu,\nu)} (\mathbb{E}_{(X,Y) \sim \gamma} \big[\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^2 \big])^{\frac{1}{2}} \tag{3}$$

where $\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^2 = \kappa(X, X) - 2\kappa(X, Y) + \kappa(Y, Y)$. If one were to purse the one-dimensional case as the easiest way to compute Wasserstein distance, the implicit feature map $\phi(\cdot)$ seems to be an obstacle. However, by the reproducing property, the expression in Eq.1 can be extended as $\langle X - Y, \omega \rangle^p = [\omega(X) - \omega(Y)]^p$ where $\omega(X)$ and $\omega(Y)$ are the witness functions. Considering $p = 2$, the squared witness functions can be expressed as $\omega(X)^2 = \langle \phi(z), \phi(X) \rangle^2$ and denoted as $\mathbb{E}[\omega^2(X)] = \langle \omega \otimes \omega, \mathbb{E}[\phi(X) \otimes \phi(X)] \rangle = \|\omega\|_\mu^2$, and similarly $\mathbb{E}[\omega^2(Y)] = \|\omega\|_\nu^2$. The max-sliced kernel Wasserstein-2 distance can be expressed in terms of squared witness functions as

$$W_2^{\mathcal{H}_*}(\mu, \nu) = \sup_{\omega \in \mathcal{H} : |\omega|_{\mathcal{H}} \leq 1} \big(\|\omega\|_\mu^2 + \|\omega\|_\nu^2 - \sup_{\gamma \in \Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{(X,Y \sim \gamma)} [2\omega(X)\omega(Y)] \big)^{\frac{1}{2}} \tag{4}$$

For a sample, the witness function $\omega(\cdot) = \sum_{i=1}^{m+n} \alpha_i \kappa(\cdot, z_i)$ is optimized in terms of a dual variables $\boldsymbol{\alpha} \in \mathbb{R}^{n+m}$ and $z_i = \left\{ \begin{array}{lll} x_i, & \text{for} & 1 \leq i \leq m \\ y_i, & \text{for} & m+1 \leq i \leq l \end{array} \right\}$, where $l = m + n$. For two finite samples, the full kernel matrix is $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{XX} & \mathbf{K}_{XY} \\ \mathbf{K}_{YX} & \mathbf{K}_{YY} \end{bmatrix}$, where $K_{i,j} = \kappa(x_i, x_i) - 2\kappa(x_i, y_j) + \kappa(y_j, y_j)$. The squared evaluations of the witness function for the sample version of the kernel max sliced

Wasserstein distance is

$$\|\omega\|_{\hat{\mu}}^2 = \sum_{j=1}^{m} \mu_j \langle \omega \otimes \omega, \phi(x_j) \otimes \phi(x_j) \rangle = \sum_{j=1}^{m} \mu_j \langle \sum_{i=1}^{n+m} \alpha_i \phi(z_i), \phi(x_j) \rangle^2 = \sum_{j=1}^{m} \mu_j \big(\sum_{i=1}^{n+m} \alpha_i \kappa(z_i, x_j) \big)^2$$
$$= \langle \boldsymbol{\mu}, \boldsymbol{\alpha}^T \mathbf{K}_{XZ}^T \mathbf{K}_{XZ} \boldsymbol{\alpha} \rangle = \boldsymbol{\alpha}^T \mathbf{K}_{XZ}^T \mathbf{D}_{\hat{\mu}} \mathbf{K}_{XZ} \boldsymbol{\alpha} = \|\mathbf{D}_{\hat{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2^2 = \frac{1}{m} \|\mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2^2, \tag{5}$$

where $\mathbf{D}_{\hat{\mu}} = \frac{1}{m}\mathbf{I}_m$, and similarly, $\|\omega\|_{\hat{\nu}}^2 = \frac{1}{n}\|\mathbf{K}_{YZ}\boldsymbol{\alpha}\|_2^2$ with $\mathbf{D}_{\hat{\nu}} = \frac{1}{n}\mathbf{I}_n$. When ω is fixed, the optimal transport plan can be obtained by sorting witness function evaluations $\omega(X) = \sum_{j=1}^{m} \langle \omega, \phi(x_j) \rangle = \mathbf{K}_{XZ}\boldsymbol{\alpha}$ and $\omega(Y) = \sum_{j=1}^{n} \langle \omega, \phi(y_j) \rangle = \mathbf{K}_{YZ}\boldsymbol{\alpha}$. The third term now can be expressed as $\sup_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y\sim\gamma)} [2\omega(X)\omega(Y)] \triangleq \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}\mathbf{K}_{YZ}\boldsymbol{\alpha}, \mathbf{K}_{XZ}\boldsymbol{\alpha} \rangle$, where $\mathcal{P}_{\hat{\mu},\hat{\nu}}$ is the transport polytope and $\mathbf{P}$ is a optimal transport plan. Kernel max-sliced Wasserstein distance can be computed in terms of witness function evaluations as

$$W_2^{\mathcal{H}_*}(\hat{\mu}, \hat{\nu})^2 = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \leq 1} \Big\{ \frac{1}{m} \|\mathbf{K}_{XZ}\boldsymbol{\alpha}\|_2^2 + \frac{1}{n} \|\mathbf{K}_{YZ}\boldsymbol{\alpha}\|_2^2 - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}\mathbf{K}_{YZ}\boldsymbol{\alpha}, \mathbf{K}_{XZ}\boldsymbol{\alpha} \rangle \Big\}. \tag{6}$$

where $\|\cdot\|_2$ is the Euclidean norm For a positive definite kernel matrix $\mathbf{K}$, the solution has to be bounded by $\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \leq 1$ in order to bound the witness function's norm.

3 Landmark-based Max-Slicing

Computation of the max-sliced Wasserstein distance still involves solving optimization problems to obtain parameters for an optimal slice which can be expensive. Here, we proposed an alternative solution called Landmark-based max-sliced Wasserstein to tackle this problem. Instead of finding the optimal slice by using complex optimization, we simply evaluate the discrepancy between distributions for each data point on Hilbert space. In the landmark case, $\boldsymbol{\alpha} \in \{e_i\}_{i=1}^l$ is the standard basis to $\mathbb{R}^l$, where $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the l-tuple with all entries are 0, except the i_{th} equals 1. Now, we can introduce landmark-based max-sliced kernel Wasserstein distance in terms of Eq.6 as following

$$W_2^{\mathcal{H}_{L^*}}(\hat{\mu}, \hat{\nu})^2 = \operatorname*{arg\,max}_{i \in \{1,2,\cdots,l\}} \Big\{ \frac{1}{m} \|\mathbf{K}_{XZ}\boldsymbol{\alpha}_i\|_2^2 + \frac{1}{n} \|\mathbf{K}_{YZ}\boldsymbol{\alpha}_i\|_2^2 - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}\mathbf{2}_{YZ}\boldsymbol{\alpha}_i, \mathbf{K}_{XZ}\boldsymbol{\alpha}_i \rangle \Big\}. \tag{7}$$

Let $K_{Xz_i} = \mathbf{K}_{XZ}\boldsymbol{\alpha}_i$ and $K_{Yz_i} = \mathbf{K}_{YZ}\boldsymbol{\alpha}_i$ represent the projections for each landmark $\{z_i\}_{i=1}^l$ from $\mathcal{Z}$. We sort the projections to max-slicing the distributions for each data point. Let $P_i, i \in [1, l]$ be the permutation which sorts the projections. The sorted projections together can be expressed as $\tilde{\mathbf{K}}_{XZ} = [\mathbf{P}_1\vec{K}_{Xz_1}, \mathbf{P}_2\vec{K}_{Xz_2}, \ldots, \mathbf{P}_l\vec{K}_{Xz_l}] = [\tilde{\mathbf{K}}_{Xz_1}, \tilde{\mathbf{K}}_{Xz_2}, \ldots, \tilde{\mathbf{K}}_{Xz_l}]$ and $\tilde{\mathbf{K}}_{YZ} = [\mathbf{P}_1\vec{K}_{Yz_1}, \mathbf{P}_2\vec{K}_{Yz_2}, \ldots, \mathbf{P}_l\vec{K}_{Yz_l}] = [\tilde{\mathbf{K}}_{Yz_1}, \tilde{\mathbf{K}}_{Yz_2}, \ldots, \tilde{\mathbf{K}}_{Yz_l}]$. For simplicity, we define Eq.7 in terms of pre-sorted witness functions as following

$$W_2^{\mathcal{H}_{L^*}}(\hat{\mu}, \hat{\nu})^2 = \operatorname*{arg\,max}_{i \in \{1,2,\cdots,l\}} \Big\{ \frac{1}{m} \|\mathbf{K}_{Xz_i}\|_2^2 + \frac{1}{n} \|\mathbf{K}_{Yz_i}\|_2^2 - 2\langle \tilde{\mathbf{K}}_{Xz_i}, \tilde{\mathbf{K}}_{Yz_i} \rangle \Big\} \tag{8}$$

Further, we reformulate the kernel max-sliced Bures distance is in terms of our proposed method. The kernel based max-sliced Bures distance in terms of squared witness functions is $\sup_{\|\omega\|_{\mathcal{H}}=1, \omega \in \mathcal{H}} = \Big| \sqrt{\mathbb{E}_{X\sim\mu}[\omega^2(X)]} - \sqrt{\mathbb{E}_{Y\sim\nu}[\omega^2(Y)]} \Big|$ The empirical sample version of Bures distance is obtained following the Eq.5. The kernel max-sliced Bures distance is $D_B^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \leq 1} \Big| \frac{1}{\sqrt{m}} \|\mathbf{K}_{XZ}\boldsymbol{\alpha}\|_2 - \frac{1}{\sqrt{n}} \|\mathbf{K}_{YZ}\boldsymbol{\alpha}\|_2 \Big|$. The landmark-based kernel max-sliced Bures distance is

$$D_B^{\mathcal{H}_{L^*}}(\hat{\mu}, \hat{\nu}) = \operatorname*{arg\,max}_{i \in \{1,2,\cdots,l\}} \Big\{ \Big| \frac{1}{\sqrt{m}} \|\mathbf{K}_{Xz_i}\|_2 - \frac{1}{\sqrt{n}} \|\mathbf{K}_{Yz_i}\|_2 \Big| \Big\} \tag{9}$$

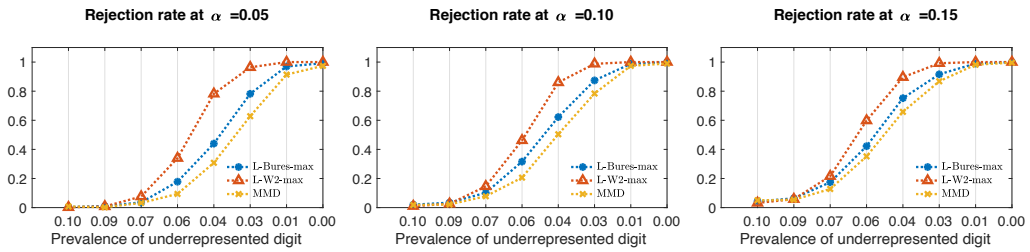

Figure 1: The figure illustrates the power test on MNIST dataset. We compare the Landmark-based kernel Bures,Wasserstein and MMD divergences. The power curves for three different critical values, α, as a function of prevalence of the underrepresented digit. For this experiment, the sample sizes are $m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

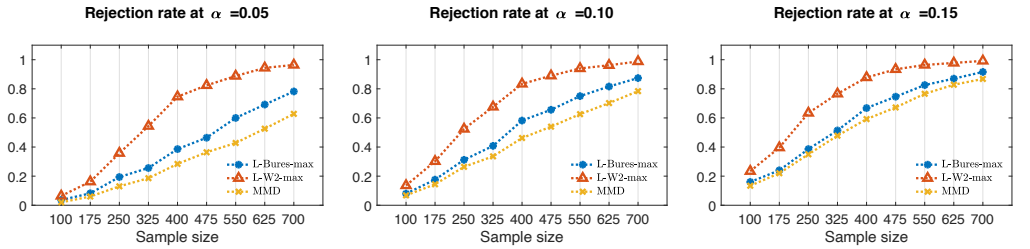

Figure 2: The figure illustrates the power test on MNIST dataset. We compare the Landmark-based kernel Bures, Wasserstein and MMD divergences. The power curves for three different critical values, α, as a function of sample sizes. For this example the prevalence of the underrepresented digit is 0.0286.

4 Covariate Shift Detection

In this section, we perform simulation experiments to compare landmark-based max-sliced divergence and MMD for detecting imbalanced classes. Specifically, we perform a statistical power test to detect the difference between a sample with a uniform distribution of classes and an imbalanced sample taken from the MNIST dataset across different levels of imbalance and sample size.

Using MNIST data-set with ten classes $l \in \{0, 1, \cdots, 9\}$, first, we test the methods when $\hat{\mu}$ is balanced but $\hat{\nu}$ is not. When uniformly sampled, each of ten digits appears on average 10% of the time. Let $p \in [0, 1]$ represent the probability that a purposefully underrepresented digit is kept when it is drawn. The probability of the underrepresented digit is calculated by $P_{under} = \frac{1}{10}(1 - p)$. 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 divergences to test the hypothesis that the two samples come from the same distribution. As a significance 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.

5 Conclusion

In this paper we have investigated the max-slicing techniques on the Wasserstein distance in Hilbert spaces to detect covariate shift. Inspired by sliced Wasserstein distance we proposed a simpler approach to solve optimal transport problems. Our approach evaluates the discrepancy between distribution for each data point on Hilbert space. Unlike the max sliced version, our method offers a simple solution. 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 covariate shift better than MMD.

References

[1] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. *On the Bures-Wasserstein distance between positive definite matrices*. 2017. arXiv: 1712.01504 [math.FA].

[2] Austin J. Brockmeier et al. *Max-sliced Bures Distance for Interpreting Discrepancies*. 2021. URL: https://openreview.net/forum?id=D2Fp_qheYu.

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

[4] Ishan Deshpande et al. *Max-Sliced Wasserstein Distance and its use for GANs*. 2019. arXiv: 1904.05877 [cs.LG].

[5] Arthur Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773. URL: http://jmlr.org/papers/v13/gretton12a.html.

[6] Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. *Sliced Wasserstein Kernels for Probability Distributions*. 2015. arXiv: 1511.03198 [cs.LG].

[7] Ker-Chau Li. "Sliced inverse regression for dimension reduction". In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[8] Cheng Meng et al. "Large-scale optimal transport map estimation using projection pursuit". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2019, pp. 8118–8129.

[9] Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. 2020. arXiv: 1803.00567 [stat.ML].

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

[11] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.

[12] Z. Zhang, M. Wang, and A. Nehorai. "Optimal Transport in Reproducing Kernel Hilbert Spaces: Theory and Applications". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.7 (2020), pp. 1741–1754. DOI: 10.1109/TPAMI.2019.2903050.