Identifying the Instances Associated with Distribution Shifts using the Max-Sliced Bures Divergence

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate an interpretable approach to compare two distributions. The approach, max-sliced Bures divergence, approximates the max-sliced Wasserstein distance, and projects the distributions into a one-dimensional subspace defined by a 'slicing' vector. Unlike heuristic algorithms for the max-sliced Wasserstein-2 distance that are not guaranteed to find the optimal slice, we detail a tractable algorithm that finds the global optimal slice and scales to large sample sizes, due to its expression in terms of second moments. However, it is unable to detect changes in higher-order statistics. To overcome this, we explore using a non-linear mapping provided by the internal representation of a pre-trained neural network (Inception Net). Our approach provides an interpretation of the Fréchet Inception distance by identifying the instances that are either overrepresented or underrepresented with respect to the other sample. We apply the proposed measure to detect class imbalances and underrepresentation within data sets.

1 Introduction

Divergence measures quantify the dissimilarity between probability distributions and are useful for detecting and correcting covariate shift [34, 28, 19]. Although a variety of divergences exist, not all are easy to both interpret and estimate. We consider a divergence to be interpretable if it can be expressed in terms of a 'witness function' that can be used to identify a measurable subset that has much higher probability under one distribution versus the other. Additionally, rather than a witness function that has non-zero evaluations throughout the domain, an interpretable divergence should identify localized subsets of the domain. This allows the user to pinpoint and understand specific examples from a subset of the differences between two samples or distributions.

A natural choice for a witness function is the log of the ratio of the probability density functions, as in the Kullback-Leibler divergence [17] and other f-divergences [1, 29]. However, estimating the densities from samples is challenging [36], which motivates alternative approaches to directly estimate the density ratio [24, 14, 38, 39, 31, 18] or variational formulations [25, 26].

Another class of divergences are the integral probability metrics (IPMs) [21]. An IPM can be expressed as the maximal difference between the first moments of the witness function evaluations under each distribution for a given function family Ω, $D_\Omega(\mu, \nu) = \sup_{\omega \in \Omega} \left| \mathbb{E}_{X \sim \mu}[\omega(X)] - \mathbb{E}_{Y \sim \nu}[\omega(Y)] \right|$. Different choices of Ω yield different IPMs; examples include total variation, the dual formulation of the Wasserstein-1 distance, maximum mean discrepancy (MMD) [11], and covariance feature matching [20]. Given $\omega^*(\cdot) \in \arg\max_{\omega \in \Omega} \mathbb{E}_{X \sim \mu}[\omega(X)] - \mathbb{E}_{Y \sim \nu}[\omega(Y)]$, one can assume that the relatively large evaluations of $\omega^*(\cdot)$ are associated to subsets with higher probability under μ. Restricting Ω to the class of functions whose large evaluations are localized would yield an interpretable divergence. An alternative option for interpretable witness functions is to consider restricted subspaces (slices) of the domain.

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

Before answering the question, 'Do existing interpretable divergences exist?' we first consider why they may not be useful for other tasks—beyond interpretability. A divergence that is restricted to localized witness functions makes inefficient use of a sample in learning and hypothesis testing (two-sample tests). For the latter, the witness function should collect evidence across the domain: restricting to a subset or subspace would limit the evidence collection. In batch sampling for generative modeling, a localized witness would only provide feedback for one specific subset at a time, wasting points that fall outside of this subset. Restricting the witness function to a rank-1 subspace would waste differences normal to this subspace.

Nonetheless, such a divergence exists: it is the max-sliced Wasserstein-p distance [16, 6], $D_{W_p,\Omega}(\mu,\nu) = \sup_{\omega\in\Omega} W_p(\omega_\sharp\mu, \omega_\sharp\nu)$ for an approrpriate function family Ω. The limitations of max-slicing for generative model and learning have already been noted and alternatives proposed: While not as simple to interpret as the general case, the recently introduced distributional form of the sliced Wasserstein distance [23], which optimizes the distribution over the slices, can also be used to investigate discrepancies across different subspaces. Another difficulty with $D_{W_p,\Omega}(\mu,\nu)$ is its computation, since it is a saddle-point optimization that requires solving a one-dimensional optimal transport problem at each evaluation. Previous works have approximated the max-sliced Wasserstein distance using a first moment approximation [6] or a finite number of steps of a local optimizer [16]. Neither guarantees obtaining an optimal witness function.

As an alternative we propose the "max-sliced Bures distance" that relaxes the Wasserstein-2 (W2) distance with a second-moment approximation that relies on the Bures distance [7, 10], a distance metric between positive semidefinite matrices, which is equivalent to the W2 distance for zero-mean Gaussians [22, 40, 27, 4, 13].

- It provides an interpretable witness functions that highlight regions of over- and under-representation in one distribution with respect to the other, especially useful for covariate shift detection.
- It is expressed as an optimization problem that has a globally optimal solution, which can be computed efficiently.
- After initial computation of the means and covariance matrices, calculating the divergence and the optimal witness function is independent of the sample sizes, enabling it to be applied to large data sets and high-dimensional settings.
- It can be easily extended beyond second-order moments by using a nonlinear mapping, for example the hidden-layer activations of a pre-trained neural network can be used to create a task-relevant mapping.

2 Methodology

We consider the set of probability distributions over the domain $\mathcal{X}$ and denote them $\mathcal{P}_\mathcal{X}$. Let $\mu,\nu \in \mathcal{P}_\mathcal{X}$ denote two probability measures, and let $X \sim \mu$ and $Y \sim \nu$ denote two random variables $X, Y \in \mathcal{X}$. We restrict our attention to probability distributions in the d-dimensional real-valued vector space $\mathcal{X} \subseteq \mathbb{R}^d$ with finite second moments and denote them as $\mathcal{P}_{L_2(\mathcal{X})}$. Denote by $\mathbf{m}_X = \mathbb{E}_{X\sim\mu}[X] \in \mathcal{X}$ and $\mathbf{m}_Y = \mathbb{E}_{Y\sim\nu}[Y] \in \mathcal{X}$ the first moments of the random variables. The family of probability mass functions over the discrete set $[n] = \{1,\dots,n\}$ is denoted $\mathcal{P}_{[n]}$. Two samples from μ and ν, $\{x_i = \mathbf{x}_i\}_{i=1}^m$ and $\{y_i = \mathbf{y}_i\}_{i=1}^n$, are denoted by their empirical measures $\hat{\mu}$ and $\hat{\nu}$. The max-sliced Bures divergence requires the matrices $\boldsymbol{\rho}_X$ and $\boldsymbol{\rho}_Y$, which can be estimated from sample means $\mathbf{m}_X, \mathbf{m}_Y$ and covariance matrices $\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y$: $\boldsymbol{\rho}_X = \mathbf{m}_X\mathbf{m}_X^\top + \frac{m-1}{m}\boldsymbol{\Sigma}_X$ and $\boldsymbol{\rho}_Y = \mathbf{m}_Y\mathbf{m}_Y^\top + \frac{n-1}{n}\boldsymbol{\Sigma}_Y$, assuming unbiased covariance estimates.

Let $\Omega = \{\omega(\cdot) = \langle\cdot,\mathbf{w}\rangle^2 : \mathbf{w}\in\mathcal{S}\}$, where $\mathcal{S}$ denotes the unit hypersphere $\mathbb{S}^{d-1}$. Using $\mathbb{E}[\omega(X)] = \mathbb{E}[\langle X,\mathbf{w}\rangle^2] = \mathbf{w}^\top\mathbb{E}[XX^\top]\mathbf{w} = \mathbf{w}^\top\boldsymbol{\rho}_X\mathbf{w}$, the max-sliced Bures divergence (MSB) is

$$D_{\text{MSB}}(\mu,\nu) = \sup_{\omega\in\Omega}|\sqrt{\mathbb{E}[\omega(X)]} - \sqrt{\mathbb{E}[\omega(Y)]}| = \sup_{\mathbf{w}\in\mathcal{S}}\left|\sqrt{\mathbf{w}^\top\boldsymbol{\rho}_X\mathbf{w}} - \sqrt{\mathbf{w}^\top\boldsymbol{\rho}_Y\mathbf{w}}\right| \tag{1}$$

$$= \max\left\{\sqrt{\mathbb{E}[\omega_{\mu>\nu}(X)]} - \sqrt{\mathbb{E}[\omega_{\mu>\nu}(Y)]}, \sqrt{\mathbb{E}[\omega_{\mu<\nu}(Y)]} - \sqrt{\mathbb{E}[\omega_{\mu<\nu}(X)]}\right\}, \tag{2}$$

where $\omega_{\mu>\nu}(\cdot) = \langle\cdot,\mathbf{w}_{\mu>\nu}\rangle^2$ and $\omega_{\mu<\nu}(\cdot) = \langle\cdot,\mathbf{w}_{\mu<\nu}\rangle^2$ correspond to the optimal slices $\mathbf{w}_{\mu>\nu}$ and $\mathbf{w}_{\mu<\nu}$. In Algorithm 1, we describe how to compute the vector $\mathbf{w}_{\mu>\nu}$ defining the one-sided

max sliced Bures divergence witness function $\omega_{\mu>\nu}$. It requires solving a one-dimensional bounded optimization, in which each function evaluation requires solving a maximum eigenvalue problems, which can be done efficiently using standard linear algebra packages. As an alternative, first-order

Algorithm 1: One-sided max-sliced Bures divergence

Input: $\boldsymbol{\rho}_X = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top, \boldsymbol{\rho}_Y = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top \in \mathbb{R}^{d\times d}, \epsilon > 0$

Define the function $\mathbf{v}_{(\cdot)} : \mathbb{R} \to \mathbb{R}^d$ as $\mathbf{v}_{(\gamma)} : \gamma \mapsto \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2\leq 1} \mathbf{w}^\top(\gamma\boldsymbol{\rho}_X - \boldsymbol{\rho}_Y)\mathbf{w}$

Solve the 1D bound problem: $\gamma^\star = \arg\max_{0<\gamma\leq 1} \sqrt{\mathbf{v}_{(\gamma)}^\top\boldsymbol{\rho}_X\mathbf{v}_{(\gamma)}} - \sqrt{\mathbf{v}_{(\gamma)}^\top\boldsymbol{\rho}_Y\mathbf{v}_{(\gamma)}}$

Output: $\mathbf{w}_{\mu>\nu} = \mathbf{v}_{(\gamma^\star)}$

constrained algorithms can be applied after the objective of the maximization is made differentiable by smoothing the negative square root $\mathbf{w}_{\mu>\nu} \approx \arg\max_{\mathbf{w}\in\mathcal{S}} \sqrt{\mathbf{w}^\top\boldsymbol{\rho}_X\mathbf{w}} - \sqrt{\mathbf{w}^\top\boldsymbol{\rho}_Y\mathbf{w} + \epsilon}$,

3 Experiments

We present various examples of using the proposed max-sliced divergences to identify the discrepancies between two samples. For illustration, we use the MNIST, CIFAR10, and Office data sets, creating pseudo-real experiments by taking different subsets of real data sets. We also use the witness function to interpret samples from GANs trained stacked MNIST.

3.1 Detecting Non-Uniform Labels Distributions

We consider covariate shift on CIFAR10, created through purposeful class imbalance. $\hat{\mu}$ is a uniform sample from the training set and $\hat{\nu}$ is a sample from the test set with less instances from one class. Applying the max-sliced Bures witness function $\omega_{\mu>\nu}$ to points from $\hat{\nu}$ should assign larger values to points from the under-represented class, identifying the top-10 witness points with the largest witness function evaluations $\omega_{\mu>\nu}(y_{\grave{\sigma}(1)}) \geq \ldots \geq \omega_{\mu>\nu}(y_{\grave{\sigma}(10)})$, where $\grave{\sigma}$ is the required permutation.

For the vector representations of the samples, we use the internal representation of the Inception Network [35] as in the Fréchet Inception distance [12] and Inception score [32]. The witness function is $\omega_{\mu>\nu}(x) = \langle \mathbf{w}_{\mu>\nu}, \phi(x)\rangle^2$, where $\phi(x) \in \mathbb{R}^{2048}$ is the Inception code. The results are reported in Figure 1. When the class probabilities differ by 2% (10.2% for majority and 8.2% for minority), the

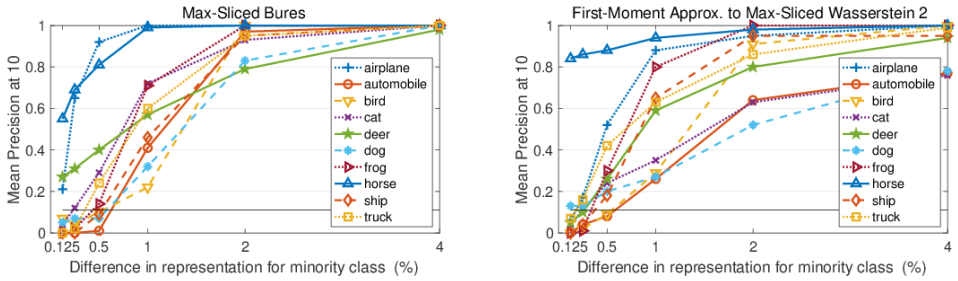

Figure 1: Max-sliced Bures distance using the Inception Network representation are applied to samples with mismatched class distributions in CIFAR10. The first sample $\hat{\mu}$ consists of the training set (balanced classes with m=50,000), and the second sample $\hat{\nu}$ is an imbalanced subset of the test set with n=10,000, wherein the noted class is less prevalent. Each curve is the mean precision@10 (mean taken across 10 random draws) for test sets with the class noted in the legend being the minority class.

mean average precision (MAP) is 0.94. This is compared to a MAP of 0.82 for the first-moment-based surrogate of the max-sliced W2 distance [6].

3.2 Detecting Mode Dropping

We use the one-sided max-sliced Bures divergence to monitor mode dropping during GAN training for the stacked MNIST data set. We create a 3-channel "Stacked MNIST" data set with 500,000

images from the MNIST training set to test mode dropping detection throughout GAN training (DCGAN architecture). The training set $\hat{\mu} = \text{Real}$ has 1000 possible modes corresponding to all 3-digit combinations. At the end of each training epoch (1000 iterations), a sample $\hat{\nu} = \text{Fake}$ of $n = 10^4$ synthetic images is generated. Figure 3 shows the top-5 fake and real images for both both witness functions after 10 epochs. To verify mode coverage we use a 4-layer convolutional network trained on single-channel MNIST; any missing combination of 3-digit labels is considered a dropped mode. Figure 2 shows the precision of the approach in detecting the dropped modes.

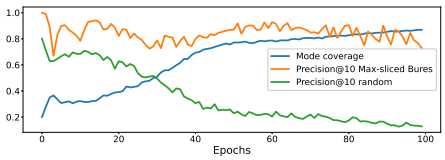

Figure 2: Detecting dropped modes using one-sided max-sliced Bures distance. The slice $\mathbf{w}_{\text{Real}>\text{Fake}}$ is used to identify the top-10 *real* training images with the largest witness function values for the slice $\mathbf{w}_{\text{Real}>\text{Fake}}$ at each epoch. Precision@10 measures the fraction that correspond to dropped modes, as compared to random selection. The curves are the mean across 100 training trials of different GAN instances.

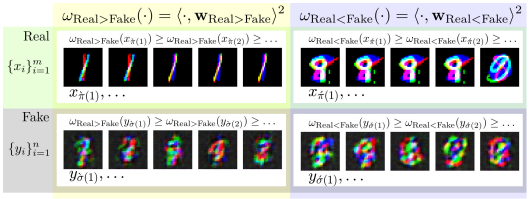

Figure 3: Example images identified by the one-sided max-sliced Bures distance. The real images correspond to a dropped mode 1-1-1 (red-green-blue), as witnessed by the closest synthetic images which are not from the same mode. The fake images for slice $\mathbf{w}_{\text{Real}<\text{Fake}}$ highlight a subset of poorly formed synthetic images that are closest to the real mode 8-9-9.

3.3 Interpreting the Fréchet Inception Distance

We use a witness function to identify identify mismatches in domain transfer tasks, specifically the Office dataset introduced by Saenko et al. [30]. The data set contains three domains: images from Amazon (A), high-resolution DSLR images (D), and low-resolution webcam images (W). There are 31 classes and a varying number of images per class. The images are represented by internal activations of the Inception object classifying network [35]. We applied both one-sided max-sliced Bures divergences to identify the two witness functions. Figure 4 shows the images identified by the witness functions. The results make sense as bicycles and notebooks without backgrounds are much more common in A than in W. The closest examples are bicycles in front of white walls. Likewise binders with a wooden table and other background are common in webcam images but uncommon on Amazon.

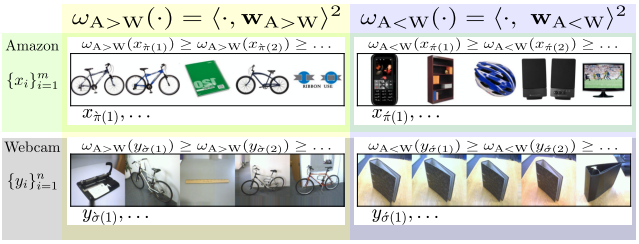

Figure 4: Max-slicing Inception codes to illustrate discrepancies in domain transfer: images from Amazon.com (A) compared to webcam images (W).

4 Conclusion

We propose the max-sliced Bures distance, a lower-bound on the max-sliced W2 distance. The underlying one-sided max-sliced Bures divergences enable direct interpretation of under- and over-representation between two samples via the witness functions, which can be used to identify discrepancies. We show the potential of applying this approach to covariate shift detection.

References

[1] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[2] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[3] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[4] Hannes De Meulemeester, Joachim Schreurs, Michaël Fanuel, Bart De Moor, and Johan AK Suykens. The bures metric for taming mode collapse in generative adversarial networks. *arXiv preprint arXiv:2006.09096*, 2020.

[5] Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

[6] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

[7] D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[8] Maurice Fréchet. Sur la distance de deux lois de probabilité. *C. R. Math. Acad. Sci. Paris*, 244 (6):689–692, 1957.

[9] Christopher Fuchs and Jeroen Van De Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *Information Theory, IEEE Transactions on*, 45(4):1216–1227, 1999.

[10] Matthias Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

[11] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, Cambridge, MA, 2007.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[13] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33, 2020.

[14] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

[15] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436, 2018.

[16] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019.

[17] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[18] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019.

[19] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130, 2018.

[20] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. McGan: Mean and covariance feature matching GAN. In *International Conference on Machine Learning*, pages 2527–2535, 2017.

[21] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

[22] Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 10237–10248. Curran Associates, Inc., 2018.

[23] 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`.

[24] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in Neural Information Processing Systems*, pages 1089–1096, 2008.

[25] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[26] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc., 2016.

[27] Jung Hun Oh, Maryam Pouryahya, Aditi Iyer, Aditya P Apte, Joseph O Deasy, and Allen Tannenbaum. A novel kernel Wasserstein distance on Gaussian measures: An application of identifying dental artifacts in head and neck computed tomography. *Computers in Biology and Medicine*, page 103731, 2020.

[28] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[29] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[30] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[31] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[33] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94, 2015.

[34] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[36] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2013.

[37] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3713–3722, 2019.

[38] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems*, pages 594–602, 2011.

[39] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25 (5):1324–1370, 2013.

[40] Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel Hilbert spaces: theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1741–1754, 2020.

A Appendix

First, we detail the relationship between Wasserstein-2 distance the Fréchet distance/divergence and the Bures distance/divergence. Then we detail additional synthetic results.

A.1 Bures, Fréchet, and Wasserstein Distances

The Bures distance generalizes the Hellinger distance to positive semidefinite matrices [9, 2]. For a pair of positive semidefinite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ the Bures distance is defined as

$$d_{\mathrm{B}}(\mathbf{A}, \mathbf{B}) \triangleq \sqrt{\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1 - 2\left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\right\|_1}, \tag{3}$$

where $\|\cdot\|_1$ denotes the trace norm (sum of the singular values), $\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\|_1 = \mathrm{tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} = \mathrm{tr}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{V}^{\frac{1}{2}})^{\frac{1}{2}}$. $d_{\mathrm{B}}(\mathbf{A}, \mathbf{A}) \leq \|\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}}\|_2^2 = \mathrm{tr}(\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}})^2$, with equality if the matrices commute. When the matrices commute they share the same eigenbasis and the Bures distance is proportional to the Hellinger distance between the eigenvalues. While the Bures distance is a metric between positive semidefinite matrices, the corresponding divergence,

$$D_{\mathrm{B}}(\mu, \nu) \triangleq d_{\mathrm{B}}(\underbrace{\mathbb{E}_{X\sim\mu}[XX^\top]}_{\boldsymbol{\rho}_X = \mathbf{m}_X\mathbf{m}_X^\top + \boldsymbol{\Sigma}_X}, \underbrace{\mathbb{E}_{Y\sim\nu}[YY^\top]}_{\boldsymbol{\rho}_Y = \mathbf{m}_Y\mathbf{m}_Y^\top + \boldsymbol{\Sigma}_Y}), \tag{4}$$

is a semi-metric since it relies on the uncentered second moments. Thus, it is a probability metric only on the subset of distributions defined completely by their first or second moments, such as Gaussian distributions or other elliptically symmetric distributions.

To compare two Gaussian distributions, it is more natural to compare the means and covariance parameters as in the Fréchet distance [8, 7],

$$D_{\mathrm{F}}(\mu, \nu) = \sqrt{\|\mathbf{m}_X - \mathbf{m}_Y\|_2^2 + d_{\mathrm{B}}^2(\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y)}. \tag{5}$$

The Fréchet distance is also known as the Wasserstein-Bures [13] or Bures-Wasserstein distance [2], since it is equal to the Wasserstein-2 distance (W_2) between Gaussian measures

$$D_{\mathrm{F}}(\mu, \nu) = W_2\left(\mathcal{N}(\mathbf{m}_X, \boldsymbol{\Sigma}_X), \mathcal{N}(\mathbf{m}_Y, \boldsymbol{\Sigma}_Y)\right). \tag{6}$$

In general, the Fréchet distance lower bounds the W2 distance [10], $D_{\mathrm{F}}(\mu, \nu) \leq W_2(\mu, \nu)$, where

$$W_2(\mu, \nu) \triangleq \sqrt{\inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|_2^2}, \tag{7}$$

and $\Gamma(\mu, \nu)$ defines the set of all joint distributions with marginals μ and ν.

The sliced Wasserstein distance [37, 5, 15] and the max-sliced Wasserstein distance [6], and its generalizations [16], evaluate discrepancies along one-dimensional linear, or non-linear, subspaces. They can be expressed by the Radon transform of integrable functions [3]. The linear max-sliced Wasserstein-2 distance (squared) is

$$D^2_{\text{MSW}_2}(\mu, \nu) = \sup_{\mathbf{w} \in \mathcal{S}} \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}[\langle X - Y, \mathbf{w}\rangle^2],$$

where $\mathcal{S}$ can be either the unit hypersphere $\mathbb{S}^{d-1}$ or its convex hull $\{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \leq 1\}$ (in terms of the optimization problem the two are equivalent as the optimal solution necessarily lies on the unit hypersphere). Slicing is motivated by the relative ease of computing the Wasserstein-p distance in one dimension, since it has a closed form [33, Lemma 2.8] and in the case of samples only requires a sorting procedure.

While the Bures distance already has a closed-form solution, the Bures distance is not easily interpretable as it cannot be written in terms of witness function. We propose a sliced version based on projecting a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ to a rank-1 matrix $\mathbf{\Omega_w} = \mathbf{w}\mathbf{w}^\top$. For compactness we denote the rank-1 projection as

$$\mathbf{A_w} = \langle \mathbf{A}, \mathbf{\Omega_w}\rangle \mathbf{\Omega_w} = \langle \mathbf{A}, \mathbf{w}\mathbf{w}^\top\rangle(\mathbf{w}\mathbf{w}^\top). \tag{8}$$

When $\mathbf{w} \in \mathbb{S}^{d-1}$, $\|\mathbf{w}\|_2 = \text{tr}(\mathbf{w}\mathbf{w}^\top) = 1$, $\text{tr}(\mathbf{A_w}) = \langle \mathbf{A}, \mathbf{w}\mathbf{w}^\top\rangle = \mathbf{w}^\top\mathbf{A}\mathbf{w}$, and

$$\begin{aligned} d_{\text{B}}(\mathbf{A_w}, \mathbf{B_w}) &= \sqrt{\mathbf{w}^\top(\mathbf{A}+\mathbf{B})\mathbf{w} - 2\sqrt{\mathbf{w}^\top\mathbf{A}\mathbf{w}}\sqrt{\mathbf{w}^\top\mathbf{B}\mathbf{w}}} \\ &= |\sqrt{\mathbf{w}^\top\mathbf{A}\mathbf{w}} - \sqrt{\mathbf{w}^\top\mathbf{B}\mathbf{w}}|. \end{aligned}$$

Let $\boldsymbol{\rho}_{X,\mathbf{w}} = \langle \mathbf{\Omega_w}, \boldsymbol{\rho}_X\rangle \mathbf{\Omega_w}$ and $\boldsymbol{\rho}_{Y,\mathbf{w}} = \langle \mathbf{\Omega_w}, \boldsymbol{\rho}_Y\rangle \mathbf{\Omega_w}$. The max-sliced Bures divergence (MSB) is

$$\begin{aligned} D_{\text{MSB}}(\mu, \nu) &\triangleq \sup_{\mathbf{w}\in\mathcal{S}} d_{\text{B}}(\boldsymbol{\rho}_{X,\mathbf{w}}, \boldsymbol{\rho}_{Y,\mathbf{w}}) \\ &= \sup_{\mathbf{w}\in\mathcal{S}} \left|\sqrt{\mathbf{w}^\top\boldsymbol{\rho}_X\mathbf{w}} - \sqrt{\mathbf{w}^\top\boldsymbol{\rho}_Y\mathbf{w}}\right| \\ &= \sup_{\mathbf{w}\in\mathcal{S}} \left|\sqrt{\mathbb{E}[\langle X, \mathbf{w}\rangle^2]} - \sqrt{\mathbb{E}[\langle Y, \mathbf{w}\rangle^2]}\right| \end{aligned}$$

A.2 Synthetic Data

We start by comparing the proposed max-sliced Bures distance to the max-sliced W2 distance for two-dimensional data. We compare Algorithm 1 for solving each one-sided max-sliced Bures divergence with gradient-based approaches for it and the max-sliced W2 distance using ADAM with parameters `lr = 1e-3, beta1 = 0.9, beta2 = 0.999, epsilon = 1e-08`, capping the number of iterations at 1000 or until the change in the slice is minimal $\|\mathbf{w} - \mathbf{w}_{\text{old}}\|_\infty < 10^{-6}$.

In two-dimensions, a near optimal slice can be obtained by a fine grid search of the sliced Bures and sliced W2 distance as shown in Figure 5 for two zero-mean Gaussian distributions.

A.3 Pseudo-real Data: Detecting Non-Uniform Labels Distributions

We investigate the precision of the witness functions in detecting covariate shift caused by class imbalances. $\hat{\mu}$ is a uniform sample from the training set and $\hat{\nu}$ is a sample from the test set with less instances from one class, the minority class. Using the one-sided max-sliced Bures divergence we obtain the optimal witness function $\omega_{\mu>\nu}$, which should have larger evaluations for at least a subset of points in the under-represented class.

We can use this witness function to identify the minority class in the test set by examining the training set instances and their labels with the largest witness function evaluations. We apply $\omega_{\mu>\nu}$ to the sample $\hat{\mu}$, identifying the top-10 witness points with the largest magnitude witness function evaluations $\omega_{\mu>\nu}(x_{\grave{\pi}(1)}) \geq \ldots \geq \omega_{\mu>\nu}(x_{\grave{\pi}(10)})$]. $\{x_{\grave{\pi}(i)}\}_{i=1}^K$ are the K points from $\hat{\mu}$ with the largest norm after projection to this subspace defined by $\mathbf{w}_{\mu>\nu}$. The performance is quantified by the precision of the labels of these instances. Ideally, these points should all be from the minority class.

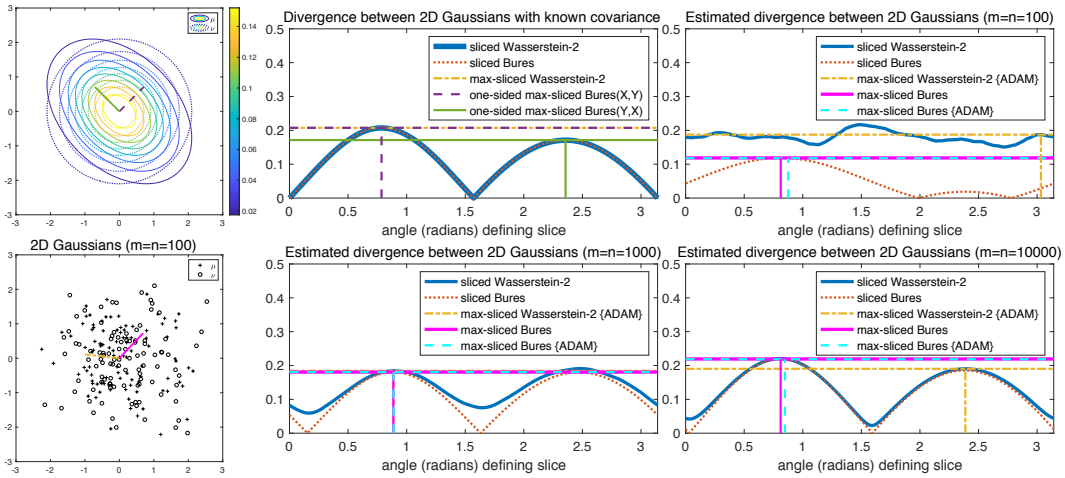

Figure 5: Sliced and max-sliced Bures and Wasserstein-2 distances are compared on population statistics and samples of varying sizes. $\mu = \mathcal{N}(\mathbf{0}, \mathbf{C})$ and $\nu = \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{C} = \mathbf{Z}\mathbf{Z}^\top$, and $\mathbf{Z} \in \mathbb{R}^{2\times 2}$ with entries that are originally standard normals and then row normalized such that $\mathbf{C}$ is a correlation matrix. In the population case and for zero-mean Gaussians, the Bures distance is equivalent to the W2 distance [10]. In the sample case, it is a lower bound. At both $m = 100$ and $m = 10^4$ the gradient optimization of the max-sliced W2 distance fails to obtain the global optimal slice (instead obtaining a local optimum).

We illustrate this simple case of covariate shift detection on the MNIST data set. Figure A.3 details a comparison of the max-sliced versions of the Bures, Fréchet, and W2 distances in terms of the divergence value estimates, the runtime, and the witness function's precision.

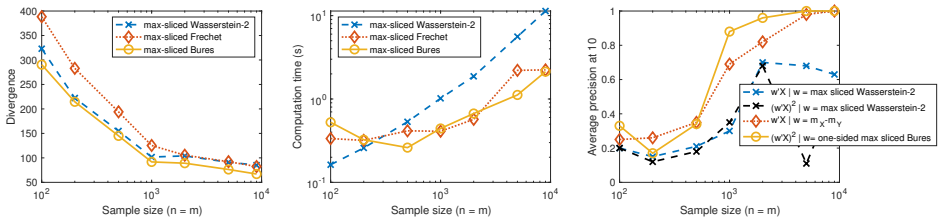

Figure 6: Max-sliced distances applied to two samples from MNIST. The first sample $\hat{\mu}$ consists of m images drawn uniformly from the training set, such that each digit has prevalence of 10%, and the second sample $\hat{\nu}$ is n images from the test set where one digit is a minority class $l \in \{0, \ldots, 9\}$ with prevalence of 5%. (Left) Divergence estimates across sample size with $l = 7$. For $m < 2000$, gradient-based approach (ADAM) for the max-sliced W2 distance fails to obtain the optimal slice as it should upper bound the max-sliced Fréchet distance. (Center) Computation time. (Right) Each curve is the average precision@10 (averaged across the 10 classes). The one-sided max-sliced Bures yields the witness function $\omega_{\mu>\nu}(\cdot) = \langle \mathbf{w}, \cdot \rangle^2$, which is applied to $\hat{\mu}$ and reliably identifies points from the minority class for $m \geq 1000$. In this case the witness function corresponding to the difference of means $\omega(\cdot) = \langle \mathbf{w}, \cdot \rangle|_{\mathbf{w}=\mathbf{m}_X - \mathbf{m}_Y}$ also works, but with lower precision for $m \geq 1000$.