MAX-SLICED BURES DISTANCE FOR INTERPRETING DISCREPANCIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the max-sliced Bures distance, a lower bound on the max-sliced Wasserstein-2 distance, to identify the instances associated with the maximum discrepancy between two samples. By scoring instances by an energy-based witness function, the proposed divergence can identify the specific subsets that are not well matched between the distributions. Localizing these discrepancies can be used to detect and correct for covariate shift and to evaluate generative adversarial networks. Similar to the Fréchet distance, the max-sliced Bures distance can be computed using the first and second-order moments and is able to capture changes in higher-order statistics through a non-linear mapping. We explore two types of non-linear mappings: positive semidefinite kernels where the witness functions belong to a reproducing kernel Hilbert space, and task-relevant mappings learned by a neural network. In the context of samples of natural images, our approach provides an interpretation of the Fréchet inception distance by identifying the synthetic and natural instances that are either over-represented or under-represented with respect to the other sample. We apply the proposed measure to detect imbalances in class distributions in various data sets and to critique generative models.

1 INTRODUCTION

Divergence measures quantify the dissimilarity or distance between distributions. They are fundamental to hypothesis testing and the estimation and criticism of statistical models, and for the optimization of generative adversarial neural networks (GANs). Although a multitude of divergences exists, not all of them are interpretable. A divergence is interpretable if it identifies the specific subsets that are not well matched between the distributions. Instances in these subsets are not outliers, but rather they reflect over- or under-representation in the other sample. Localizing these discrepancies is incredibly useful for understanding and compensating for differences between two samples or distributions, to detect covariate shift (Shimodaira, 2000; Quionero-Candela et al., 2009; Lipton et al., 2018) or to evaluate generative models (Heusel et al., 2017). We propose a novel divergence measure, the max-sliced Bures distance, to identify the instances associated with the maximum discrepancy between two samples through the use of a *witness function* that maps instances in the sample space to a real-valued score. The divergence is simply the difference between the moments of the optimal witness function evaluations for each distribution. The optimal witness function is selected from a suitable family to maximize this difference. The witness function identifies subsets of instances within one sample that are poorly represented in the other sample.

While many divergences can be posed in terms of witness functions (see Table 1), not all witness functions are readily obtained. From an information-theoretic perspective, the most natural witness function is the logarithm of the ratio of the densities (Kullback & Leibler, 1951) as in the Kullback-Leibler divergence. Applying various other convex functions to the density ratio constitutes the family of f-divergences (Ali & Silvey, 1966; Rényi, 1961), which include the Hellinger, Jensen-Shannon, Pearson χ^2, and others. However, without a parametric model, estimating the densities, and thus the witness functions, from samples is challenging (Vapnik, 2013). Following Vapnik's advice to "try to avoid solving a more general problem as an intermediate step," previous work has sought to directly model the density ratio via a kernel-learning framework (Nguyen et al., 2008; Kanamori et al., 2009; Yamada et al., 2011; 2013; Saito et al., 2018; Lee et al., 2019) or to estimate an f-divergence directly by optimizing a function from a suitable family (Nguyen et al., 2010), for instance a neural network, as proposed by Nowozin et al. (2016).

Witness functions need not rely on the density-ratio or their approximation. A wide class of divergences called integral probability metrics (IPM) (Müller, 1997), which include the total variation, the Wasserstein-1 distance, maximum mean discrepancy (MMD) (Gretton et al., 2007), and others (Mroueh et al., 2017), seek a witness function that maximizes the distance between the first moments of the witness function evaluations. A class of related divergences are the max-sliced Wasserstein-p distances, which seeks a linear (Deshpande et al., 2019) or non-linear function (Kolouri et al., 2019), that maximizes the Wasserstein-p distance between the witness evaluations for the two distributions. The optimal function, whether linear or not, is a witness function.

We propose a novel IPM-like divergence measure that approximates the max-sliced Wasserstein-2 distance and relate it to other symmetric divergences: the total variation distance, the Hellinger distance, and the Fréchet distance. Like the Fréchet distance, it can be computed using the first and second-order moments of the data. Consequently, to detect higher-order changes in distributions the random variables must be mapped through a non-linear function. We explore two options. Firstly, using positive semidefinite kernels we formulate the divergence measure and its underlying witness function in a reproducing kernel Hilbert space (RKHS), where each point in the space is a possible witness function. The proposed divergence relies on distance measures between covariance operators, specifically the Bures distance (Bures, 1969; Uhlmann, 1976; Nielsen & Chuang, 2000; Koltchinskii & Xia, 2015), which generalizes the Hellinger distance to matrices and operators. We show how projecting these operators to rank-1 subspaces enables direct computation with input-space functionals. Secondly, we use a neural network to create a task-relevant mapping, computing the second-order statistics of the hidden-layer activations, and apply this in the context of samples of natural images. The max-sliced Bures distance provide a lower bound on the max-sliced Fréchet distance. This enables interpretation of the Fréchet inception distance (FID) (Heusel et al., 2017) by identifying the images associated with discrepancies.

For the kernel case, we detail the optimizations necessary to compute the max-sliced version of the kernel Wasserstein distance (Zhang et al., 2020), which like the other sliced Wasserstein versions (Kolouri et al., 2019) can be computed for empirical samples by sorting scalars. In contrast, the optimization required for the max-sliced Bures distance avoids the saddlepoint optimization of the max-sliced Wasserstein and outperforms on the first-order surrogate approximation of the max-sliced Wasserstein-2 distance (Deshpande et al., 2019). The max-sliced Bures is optimized as the maximum of one-sided asymmetric divergences, which aids interpretation as shown in Figure 1. Each one-sided divergence is a difference of convex functions for which tractable algorithms exist.

We apply the linear and kernel-based versions to detect imbalances in class distributions of natural images and to critique GANs. We compare to other divergences expressed in terms of witness functions including MMD and the covariance feature matching proposed by Mroueh et al. (2017).

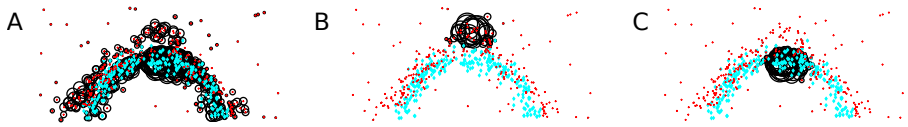

Figure 1: The magnitude of witness function evaluations (proportional to circle radius) represents the discrepancy between the two samples. In these cases, a Gaussian kernel is used to construct the witness function through (A) MMD, and (B,C) the two one-sided max-sliced kernel Bures divergence, which are much more localized—identifying only the largest discrepancies.

2 METHODOLOGY

We consider a topological space $\mathcal{X}$ with a positive-definite, bounded kernel function $\kappa : \mathcal{X} \times \mathcal{X} \to B \subset \mathbb{R}$, a Borel σ-algebra $\mathcal{B}_\mathcal{X}$, and the set $\Pr(\mathcal{X})$ of Borel probability measures on $\mathcal{X}$. For any κ, there is an implicit mapping $\phi : \mathcal{X} \to \mathcal{H}$ that maps any element $x \in \mathcal{X}$ to an element in the Hilbert space $\phi(x) \in \mathcal{H}$ such that $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle = \langle \kappa(\cdot, x), \kappa(\cdot, y) \rangle$ for $x, y \in \mathcal{X}$, and $\forall h \in \mathcal{H}, h(x) = \langle h, \kappa(\cdot, x) \rangle$ (Aronszajn, 1950). For an operator $T \in \mathcal{H} \times \mathcal{H}$ the Schatten p-norm of T is denoted $\|T\|_p$. The case of $p = 1$ corresponds to the trace norm $\|T\|_1 = \mathrm{tr}[\sqrt{T^* T}]$, where T^* denotes the adjoint of T such that $\langle \omega, T^* \psi \rangle = \langle T\omega, \psi \rangle$, for all $\omega, \psi \in \mathcal{H}$, the case $p = 2$ corresponds to the Hilbert-Schmidt norm $\|T\|_2 = \sqrt{\mathrm{tr}[T^* T]} = \sqrt{\langle T, T \rangle}$, and the case of $p \to \infty$

corresponds to the operator norm $\|T\|_\infty$. A rank-1 operator can be formed using the tensor product $h_1 \otimes h_2 \in \mathcal{H} \times \mathcal{H}$ with $\langle (h_1 \otimes h_2)\omega, \psi \rangle = \langle h_2, \omega \rangle \langle h_1, \psi \rangle$.

Let $\mu, \nu \in \Pr(\mathcal{X})$ denote two probability measures, and $X \sim \mu$ and $Y \sim \nu$ be two random variables $X, Y \in \mathcal{X}$. Denote by $m_X = \mathbb{E}_{X\sim\mu}[\phi(X)] \in \mathcal{H}$ and $m_Y = \mathbb{E}_{Y\sim\nu}[\phi(Y)] \in \mathcal{H}$ the first moments of the random variables in the RKHS. For conciseness, the measure may be dropped from the expectation when the context is clear, e.g., $m_X = \mathbb{E}[\phi(X)]$. The uncentered second moments are defined as $\rho_X = \mathbb{E}[\phi(X) \otimes \phi(X)] \in \mathcal{H} \times \mathcal{H}$ and $\rho_Y = \mathbb{E}[\phi(Y) \otimes \phi(Y)] \in \mathcal{H} \times \mathcal{H}$. As uncentered moments these operators also contain information on the first-order moment in the RKHS. The corresponding covariance operators are $\Sigma_X = \rho_X - m_X \otimes m_X$ and $\Sigma_Y = \rho_Y - m_Y \otimes m_Y$.

2.1 Divergences as Distance Metrics

Let $D(\mu, \nu)$ denote a divergence $D : \Pr(\mathcal{X}) \times \Pr(\mathcal{X}) \to [0, \infty)$. It is a distance metric between measures (a probability metric) if all of the following statements hold: (i) $\mu = \nu \implies D(\mu, \nu) = 0$, (ii) $D(\mu, \nu) = 0 \implies \mu = \nu$, (iii) $D(\mu, \nu) = D(\nu, \mu)$, (iv) $D(\mu, \nu) \leq D(\mu, \xi) + D(\nu, \xi)$. It is a semi-metric if all properties aside from (ii) hold. Müller (1997) defines the class of integral probability metrics in terms as the supremum between the absolute difference between expectations

$$D_{\mathcal{F}}(\mu, \nu) = \sup_{\omega \in \mathcal{F}} \left| \int_{\mathcal{X}} \omega(x) d\mu(x) - \int_{\mathcal{X}} \omega(x) d\nu(x) \right| = \sup_{\omega \in \mathcal{F}} \left| \mathbb{E}[\omega(X)] - \mathbb{E}[\omega(Y)] \right|.$$

With appropriate choice of the family of functions $\mathcal{F}$, this form yields well-known divergences (Sriperumbudur et al., 2010), e.g., when $\mathcal{F}$ is the set of functions with Lipschitz constant less than 1, the resulting divergence is the Wasserstein-1 distance metric. Another example is when $\mathcal{F} = \{\omega \in \mathcal{H} : \|\omega\|_2 \leq 1\}$, which yields MMD (Gretton et al., 2007), defined as

$$\begin{aligned} D^{\mathcal{H}}_{MMD}(\mu, \nu) &= \sup_{\omega \in \mathcal{H} : \|\omega\|_2 \leq 1} \{\mathbb{E}[\omega(X)] - \mathbb{E}[\omega(Y)] = \mathbb{E}[\langle \phi(X) - \phi(Y), \omega \rangle]\} = \|m_X - m_Y\|_2 \\ &= \sqrt{\mathbb{E}_{X\sim\mu, X'\sim\mu}[\kappa(X, X')] + \mathbb{E}_{Y\sim\nu, Y'\sim\nu}[\kappa(Y, Y')] - 2\mathbb{E}_{X\sim\mu, Y\sim\nu}[\kappa(X, Y)]}. \quad (1) \end{aligned}$$

For characteristic kernels such as the Laplacian and Gaussian kernels, the mean embedding $\mathbb{E}_{X\sim\mu}[\phi(X)] : \Pr(\mathcal{X}) \to \mathcal{H}$ is an injective function (Sriperumbudur et al., 2008; Fukumizu et al., 2009; Sriperumbudur et al., 2010), capturing the full statistics of μ. In these cases, MMD is a distance metric on $\Pr(\mathcal{X})$; likewise, distance metrics between the operators $\rho_X = \mathbb{E}[\phi(X) \otimes \phi(X)]$ and $\rho_Y = \mathbb{E}[\phi(Y) \otimes \phi(Y)]$ induce probability metrics for characteristic kernels (Zhang et al., 2020).

2.2 Operator Distances for Defining Divergences

Total variation (TV) is a well-known probability metric and an integral probability metric (Müller, 1997), taking the form $(1/2)\sum_i |p_i - q_i|$ for discrete measures, for which $p_i = \mu(x_i)$ and $q_i = \nu(x_i)$ where $\{x_i\}_i = \mathcal{X}$. The TV distance between operators in the RKHS is a divergence

$$D^{\mathcal{H}}_{TV}(\mu, \nu) \triangleq d_{TV}(\rho_X, \rho_Y) \triangleq \frac{1}{2}\|\rho_X - \rho_Y\|_1, \quad (2)$$

where $\|\cdot\|_1$ denotes the trace norm (Schatten 1-norm).

The Bures distance (Bures, 1969; Uhlmann, 1976) generalizes the Hellinger distance to operators, especially for comparing quantum density matrices (Fuchs & Van De Graaf, 1999; Bromley et al., 2014). Recall that the square of the Hellinger distance is $\frac{1}{2}\sum_i (\sqrt{p_i} - \sqrt{q_i})^2$ for discrete measures. The Bures distance (squared) is defined as

$$d^2_B(\rho_X, \rho_Y) \triangleq \|\rho_X\|_1 + \|\rho_Y\|_1 - 2\left\|\sqrt{\rho_X}\sqrt{\rho_Y}\right\|_1 \leq \|\sqrt{\rho_X} - \sqrt{\rho_Y}\|_2^2 = \mathrm{tr}[(\sqrt{\rho_X} - \sqrt{\rho_Y})^2], \quad (3)$$

with equality if the operators commute. The divergence is denoted $D^{\mathcal{H}}_B(\mu, \nu) = d_B(\rho_X, \rho_Y)$.

Recently, the Bures distance has been used in in the context of the Wasserstein-2 distance between Gaussian measures in an RKHS (Zhang et al., 2020). The kernel Gauss-Wasserstein distance is

$$D^{\mathcal{H}}_{GW}(\mu, \nu) \triangleq \sqrt{\|m_X - m_Y\|_2^2 + d^2_B(\Sigma_X, \Sigma_Y)} = \sqrt{[D^{\mathcal{H}}_{MMD}(\mu, \nu)]^2 + d^2_B(\Sigma_X, \Sigma_Y)}. \quad (4)$$

This is a kernel-based version of the Wasserstein-2 distance between multivariate Gaussian measures (Gelbrich, 1990), commonly known as the Fréchet distance (Fréchet, 1957; Dowson & Landau, 1982), which provides a lower bound for the Wasserstein-2 distance.[1] Zhang et al. (2020) also proposed the kernel Wasserstein-p distance between μ and ν,

$$W_p^{\mathcal{H}}(\mu,\nu) \triangleq \inf_{\gamma\in\Gamma(\mu,\nu)} \left[\mathop{\mathbb{E}}_{(X,Y)\sim\gamma} [d_\kappa^p(X,Y)] \right]^{\frac{1}{p}}, \quad p \geq 1, \tag{5}$$

where $\Gamma(\mu,\nu)$ defines the set of all joint distributions coupling μ and ν, and $d_\kappa^p(X,Y) = \|\phi(X) - \phi(Y)\|_2^p$. For $p = 2$, $d_\kappa^2(X,Y) = \kappa(X,X) + \kappa(Y,Y) - 2\kappa(X,Y)$.

2.3 Divergences based on slicing Hilbert spaces

The sliced Wasserstein distance (Wu et al., 2019; Deshpande et al., 2018; Kolouri et al., 2018), and max-sliced Wasserstein distance (Deshpande et al., 2019; Kolouri et al., 2019) evaluate discrepancies in linear or non-linear one-dimensional subspaces. A motivation for this is the analytic solution of the Wasserstein-p distance in one dimension. Similarly, we propose extensions of the kernel-based TV, Bures, and Wasserstein-p distances using the rank-1 projection operator $\Omega = \omega \otimes \omega \subset \mathcal{H} \times \mathcal{H}$, where $\|\omega\|_2 \leq 1, \omega \in \mathcal{H}$. $\Omega = \omega \otimes \omega$ projects (slices) the RKHS along a one-dimensional subspace defined by the ray ω, with $\langle \omega, \phi(X) \rangle = \omega(X)$, due to the reproducing property. In this formulation, $\omega : \mathcal{X} \to \mathbb{R}$ is the witness function. While a slice is a linear subspace in the RKHS, it corresponds to a possibly non-linear function in the input space. (In section 2.5.1, we investigate the linear case.) For conciseness, we denote the uncentered second moment or 'energy' $\langle \omega \otimes \omega, \rho_X \rangle = \mathbb{E}\left[\omega^2(X)\right]$ as $\|\omega\|_\mu^2$, and $\langle \omega \otimes \omega, \rho_Y \rangle = \mathbb{E}\left[\omega^2(Y)\right]$ as $\|\omega\|_\nu^2$. The max-sliced kernel TV and Bures distances, derived in appendix A.1, are expressed as

$$\textit{max-}D_{TV}^{\mathcal{H}}(\mu,\nu) \triangleq \frac{1}{2} \sup_{\omega:\|\omega\|_2\leq 1} \left| \|\omega\|_\mu^2 - \|\omega\|_\nu^2 \right|, \text{and} \tag{6}$$

$$\textit{max-}D_{B}^{\mathcal{H}}(\mu,\nu) \triangleq \sup_{\omega:\|\omega\|_2\leq 1} \Big| \|\omega\|_\mu - \|\omega\|_\nu \Big| = \max_{s\in\{-1,1\}} \sup_{\omega:\|\omega\|_2\leq 1} s(\|\omega\|_\mu - \|\omega\|_\nu), \tag{7}$$

respectively. For the latter, we refer to the inner supremums, $D(\mu\|\nu) = \sup_{\omega:\|\omega\|_2\leq 1} \|\omega\|_\mu - \|\omega\|_\nu$ and $D(\nu\|\mu) = \sup_{\omega:\|\omega\|_2\leq 1} \|\omega\|_\nu - \|\omega\|_\mu$, as one-sided divergences. The one-sided divergences quantify any over-representation by the first measure in terms of the difference in energy. The max-sliced TV distance is an IPM with $\mathcal{F} = \{f(x) = \omega^2(x) = \langle \phi(x), \omega \rangle^2 : \|\omega\|_2 \leq 1, \forall \omega \in \mathcal{H}\}$ and is a special case of the IPM$_\Sigma$ divergence proposed by Mroueh et al. (2017). While not an IPM, the max-sliced Bures distance uses the absolute difference of the square roots of the expectations, providing a lower-bound on the Fréchet distance between the pushforward measures $\omega_\# \mu$ and $\omega_\# \nu$. The max-sliced kernel Wasserstein-2 distance is

$$\textit{max-}W_2^{\mathcal{H}}(\mu,\nu) = \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \sqrt{\|\omega\|_\mu^2 + \|\omega\|_\nu^2 - \sup_{\gamma\in\Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{(X,Y)\sim\gamma} [2\omega(X)\omega(Y)]}, \tag{8}$$

which has a closed-form expression of the optimal γ in terms of the inverse cumulative distribution functions of $\omega_\# \mu$ and $\omega_\# \nu$ (see appendix A.1.4). We note key relationships among the divergences in appendix A.2, including showing that the max-sliced Bures distance is a lower bound on the max-sliced Wasserstein-2 distance, by way of the max-sliced Fréchet distance.

2.4 Computing the Max-sliced Divergences

The max-sliced kernel Bures, TV, and Wasserstein-2 distances require optimizations to find the optimal witness function. As noted by others investigating IPMs (Mroueh et al., 2017; Li et al., 2017; Kolouri et al., 2019), the witness function can be defined using multiple kernels or the composition of a kernel and a family of functions implemented as a neural network. This is relevant in the context of a GAN, where μ is the true distribution and ν is formed as a pushforward measure of the noise

[1] In the non-singular and finite-dimensional case, the multivariate Fréchet distance (squared) is often expressed as $\|\mathbf{m}_X - \mathbf{m}_Y\|_2^2 + \mathrm{tr}(\mathbf{\Sigma}_X + \mathbf{\Sigma}_Y - 2\sqrt{\mathbf{\Sigma}_X \mathbf{\Sigma}_Y})$; however, this expression is not accurate if either covariance is singular (Gelbrich, 1990). The trace term should be replaced by $d_B^2(\mathbf{\Sigma}_X, \mathbf{\Sigma}_Y) = \mathrm{tr}(\mathbf{\Sigma}_X + \mathbf{\Sigma}_Y) - 2\left\|\sqrt{\mathbf{\Sigma}_X}\sqrt{\mathbf{\Sigma}_Y}\right\|_1$, where $\|\mathbf{M}\|_1$ is the sum of $\mathbf{M}$'s singular values.

distribution $Z \sim \zeta$ through the generator network $G(Z) = Y \sim \nu$. In this context, Goodfellow et al. (2014) use the divergence $D_A(\mu, \nu) = \max_\omega \mathbb{E}\left[\log \omega(X)\right] + \mathbb{E}\left[\log\left(1 - \omega(Y)\right)\right]$, where $\omega : \mathcal{X} \to [0, 1]$. In comparison, using the Wasserstein-1 distance $D_{W^1}(\mu, \nu) = \sup_{\omega \in \text{Lip}^1} \mathbb{E}[\omega(X)] - \mathbb{E}[\omega(Y)]$ requires a Lipschitz constraint (Arjovsky et al., 2017; Gulrajani et al., 2017). Table 1 compares some existing divergences to the proposed[2] in terms of witness functions and constraints.

Table 1: Divergences written in terms of witness functions. Optimal solutions denoted ω^*.

$D_A(\mu, \nu)$	$= \max_{\omega : \omega(\cdot) > 0} \mathbb{E}[\log \omega(X)] + \mathbb{E}[\log(1 - \omega(Y))], \quad \omega^* = \frac{d\mu(x)}{d\mu(x) + d\nu(x)}.$
$D_{W^1}(\mu, \nu)$	$= \sup_{\omega \in \text{Lip}^1} \mathbb{E}[\omega(X)] - \mathbb{E}[\omega(Y)].$
$D^{\mathcal{H}}_{MMD}(\mu, \nu)$	$= \sup_{\omega \in \mathcal{H} : \lVert\omega\rVert_2 \leq 1} \mathbb{E}[\omega(X)] - \mathbb{E}[\omega(Y)], \quad \omega^* = \frac{m_X - m_Y}{\lVert m_X - m_Y\rVert_2}.$
max-$D^{\mathcal{H}}_{TV}(\mu, \nu)$	$= \sup_{\omega \in \mathcal{H} : \lVert\omega\rVert_2 \leq 1} (1/2) \left\lvert \mathbb{E}[\omega^2(X)] - \mathbb{E}[\omega^2(Y)] \right\rvert.$
max-$D^{\mathcal{H}}_{B}(\mu, \nu)$	$= \sup_{\omega \in \mathcal{H} : \lVert\omega\rVert_2 \leq 1} \left\lvert \sqrt{\mathbb{E}[\omega^2(X)]} - \sqrt{\mathbb{E}[\omega^2(Y)]} \right\rvert.$

2.5 Sample-Based Estimators for Max-Sliced Divergences

We consider the case of finite samples expressed as empirical measures $\hat{\mu} = \sum_{i=1}^m \mu_i \delta_{x_i}$ and $\hat{\nu} = \sum_{i=1}^n \nu_i \delta_{y_i}$ for the samples $\{x_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^n$ with discrete probability masses denoted as column vectors $[\mu_1, \ldots, \mu_m]^\top = \boldsymbol{\mu} \in [0, 1]^m$, $\langle \boldsymbol{\mu}, \mathbf{1} \rangle = 1$, and $[\nu_1, \ldots, \nu_n]^\top = \boldsymbol{\nu} \in [0, 1]^n$, $\langle \boldsymbol{\nu}, \mathbf{1} \rangle = 1$. The kernel-based max-sliced divergences optimize the witness function $\omega(\cdot) = \sum_{k=1}^{m+n} \alpha_k \kappa(\cdot, z_k)$ in terms of the dual variables $\boldsymbol{\alpha} \in \mathbb{R}^{m+n}$ corresponding to the pooled samples $z_k \in \{x_i\}_{i=1}^m \cup \{y_j\}_{j=1}^n$. The optimizations are detailed in appendix A.4. For clarity, we proceed to the linear kernel case.

2.5.1 Finite Dimensional Hilbert Space

We now consider a finite-dimensional embedding $\phi(x) = \mathbf{x} \in \mathbb{R}^d$. After embedding, kernel evaluations correspond to vector inner-products $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle = \mathbf{x}^\top \mathbf{y}$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{d \times n}$ denote the sample points with corresponding masses $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively. Evaluation of the witness function is the inner product $\omega(x) = \mathbf{w}^\top \mathbf{x}$, where the variable $\mathbf{w}$ defines the slice with $\lVert\omega\rVert^2_{\hat{\mu}} = \mathbf{w}^\top \mathbf{X} \mathbf{D}_{\boldsymbol{\mu}} \mathbf{X}^\top \mathbf{w}$ and $\lVert\omega\rVert^2_{\hat{\nu}} = \mathbf{w}^\top \mathbf{Y} \mathbf{D}_{\boldsymbol{\nu}} \mathbf{Y}^\top \mathbf{w}$, where $\mathbf{D}_{\mathbf{v}}$ is a diagonal matrix with entries $\mathbf{v}$. For i.i.d. samples, $\mathbf{D}_{\boldsymbol{\mu}} = \frac{1}{m}\mathbf{I}$ and $\mathbf{D}_{\boldsymbol{\nu}} = \frac{1}{n}\mathbf{I}$. The instances associated with a discrepancy can be identified by the witness function evaluations $\boldsymbol{\omega}_X = \mathbf{w}^\top \mathbf{X}$ and $\boldsymbol{\omega}_Y = \mathbf{w}^\top \mathbf{Y}$. The max-sliced TV, Bures, and Wasserstein-2 divergences are

$$\textit{max-}D^{\mathbb{R}^d}_{TV}(\hat{\mu}, \hat{\nu}) = \max_{\mathbf{w} : \lVert\mathbf{w}\rVert_2 \leq 1} \left\lvert \mathbf{w}^\top (\mathbf{X} \mathbf{D}_{\boldsymbol{\mu}} \mathbf{X}^\top - \mathbf{Y} \mathbf{D}_{\boldsymbol{\nu}} \mathbf{Y}^\top) \mathbf{w} \right\rvert, \tag{9}$$

$$\textit{max-}D^{\mathbb{R}^d}_{B}(\hat{\mu}, \hat{\nu}) = \max_{\mathbf{w} : \lVert\mathbf{w}\rVert_2 \leq 1} \left\lvert \sqrt{\mathbf{w}^\top \mathbf{X} \mathbf{D}_{\boldsymbol{\mu}} \mathbf{X}^\top \mathbf{w}} - \sqrt{\mathbf{w}^\top \mathbf{Y} \mathbf{D}_{\boldsymbol{\nu}} \mathbf{Y}^\top \mathbf{w}} \right\rvert, \text{ and} \tag{10}$$

$$\textit{max-}W^{\mathbb{R}^d}_{2}(\hat{\mu}, \hat{\nu}) = \max_{\mathbf{w} : \lVert\mathbf{w}\rVert_2 \leq 1} \sqrt{\mathbf{w}^\top (\mathbf{X} \mathbf{D}_{\boldsymbol{\mu}} \mathbf{X}^\top + \mathbf{Y} \mathbf{D}_{\boldsymbol{\nu}} \mathbf{Y}^\top) \mathbf{w} - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu}, \hat{\nu}}} \mathbf{w}^\top \mathbf{X} \mathbf{P}^\top \mathbf{Y} \mathbf{w}}, \tag{11}$$

where $\mathcal{P}_{\hat{\mu}, \hat{\nu}} = \{\mathbf{P} \in [0, 1]^{m \times n} | \mathbf{P} \mathbf{1}_n = \boldsymbol{\mu}, \mathbf{P}^\top \mathbf{1}_m = \boldsymbol{\nu}\}$ is a transportation polytope.

These three optimizations differ in difficulty. The first two require only the sample means $\mathbf{m}_X, \mathbf{m}_Y$ and covariance matrices $\boldsymbol{\Sigma}_X, \boldsymbol{\Sigma}_Y$, since $\mathbf{X} \mathbf{D}_{\boldsymbol{\mu}} \mathbf{X}^\top = \mathbf{m}_X \mathbf{m}_X^\top + \frac{m-1}{m} \boldsymbol{\Sigma}_X$ and $\mathbf{Y} \mathbf{D}_{\boldsymbol{\nu}} \mathbf{Y}^\top = \mathbf{m}_Y \mathbf{m}_Y^\top + \frac{n-1}{n} \boldsymbol{\Sigma}_Y$ (assuming unbiased covariance estimates). The max-sliced TV can be solved by finding the eigenvector associated to the largest magnitude eigenvalue of the difference of uncentered covariance matrices. The one-sided max-sliced Bures divergences are each a difference of convex functions (Tao & An, 1998), for which a local ascent algorithm will often yield the global optimum (details are in appendix A.4). In comparison, the max-sliced Wasserstein-2 distance is a

[2]The sum of the one-sided divergences $D(\mu \| \nu) + D(\nu \| \mu) \leq 2$*max*-$D^{\mathcal{H}}_B(\mu, \nu)$ is more natural for training a GAN, since it enables instantiation of separate witness functions for over- and under-representation. However, as it tends to localize discrepancies, the witness function may not make efficient use of the generator's samples.

saddlepoint optimization (Deshpande et al., 2019). Following Kolouri et al. (2019), gradient ascent on $\mathbf{w}$ is performed, using at each iteration, the optimal transport map $\mathbf{P}^*$ found by sorting.

3 EXPERIMENTS

We present results of using the proposed kernel and covariance-based max-sliced divergences to identify the discrepancies between two samples. In section 3.1, we apply this approach to the task of evaluating samples of natural and synthetic images using the internal representation of a pre-trained classification network, namely, the inception network (Szegedy et al., 2016) as in the Fréchet inception distance (Heusel et al., 2017) and the inception score (Salimans et al., 2016). In section 3.2, we apply max-sliced divergences to the task of detecting instances associated with mismatched distributions (covariate shift), and in section 3.3 we propose optimizing the weights $\boldsymbol{\nu}$ to compensate for covariate shift. For illustration, we use MNIST and CIFAR10 data sets.

3.1 INTERPRETING THE FRÉCHET INCEPTION DISTANCE

We use a linear witness function to identify instances that are not well matched between two samples of real or synthetic natural images represented by internal activations of the inception object classifying network (Szegedy et al., 2016). Specifically, we search for a witness function $\omega(x) = \langle \mathbf{w}, \phi(x) \rangle$ that optimizes the max-sliced divergences, where the vector $\phi(x) \in \mathbb{R}^{2048}$ is an inception code—the internal activations of penultimate layer of the network after pooling (Heusel et al., 2017).

Figure 2 shows the performance of the proposed measure to identify instances associated with imbalanced representation of particular classes. In particular, the one-sided max-sliced Bures divergence (equation 10 without the absolute value) is applied to the imbalanced sample, and the top-10 instances with the largest magnitude witness function evaluations are selected; the performance is quantified by the precision of the labels of these instances (ideally, the top instances should be from the underrepresented class). Notably, the mean precision of the top-10 instances is 0.79 or better across the classes with a mean average precision (MAP) of 0.94 when the class probabilities differ by 2% (10.2% for majority and 8.2% for minority). This is compared to a minimum of 0.52 across the classes and a MAP of 0.82 for the first-moment based surrogate of the max-sliced Wasserstein-2 distance (Deshpande et al., 2019). (The approximation of the max-sliced Wasserstein-2 distance optimization takes longer to match the same performance of the max-sliced Bures distance.)[3]

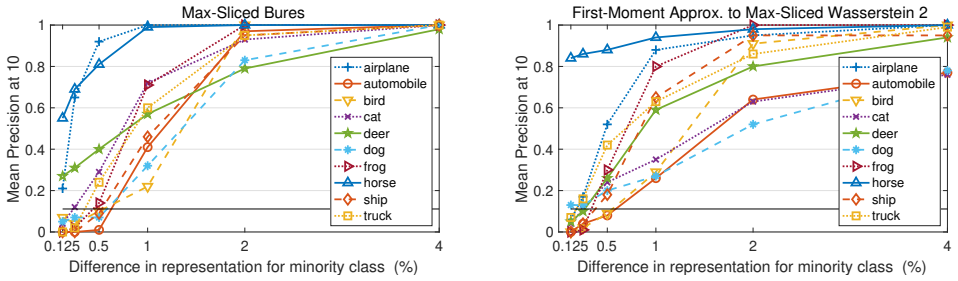

Figure 2: Max-sliced divergences using the inception network representation are applied to samples with mismatched class distributions in CIFAR10. The first sample consists of the training set (balanced classes with m=50,000), and the second sample is an imbalanced subset of the test set with n=10,000. Each curve is the mean precision@10 (averaged across 10 random draws) for test sets where the given class is subsampled at different levels of imbalance and other classes are balanced.

Next we generate a set of 50,000 synthetic images for an AutoGAN instance pre-trained on CIFAR10 (Gong et al., 2019), which has an inception score of 8.525 and a FID score of 12.41. We applied both one-sided max-sliced Bures divergences to identify the subspaces that maximizes

[3]For a matched sample size of $m = n = 2000$, the mean and standard deviation of the run time of the max-sliced Wasserstein-2 distance is 81.7±30.4s compared to 0.69±0.27s for the one-sided max-sliced Bures. Both are implement in MATLAB using a quasi-Newton optimizer (Schmidt, 2012) with default settings, running on a 2.7 GHz Intel Core i7 with 16 GB of RAM.

the difference in energy between synthetic and real images. We also computed the realism scores R (Kynkäänniemi et al., 2019) of the entire set of synthetic images using the set of 10,000 test images and 3-nearest neighbor distances. The Spearman rank correlation between the realism scores for the full set of synthetic images and the witness function evaluations is -0.70 for the witness function that has high energy for synthetic images and low energy for real images, and it is 0.17 for the other witness function. Figure 3 details the top-10 images in each subspace and their realism scores. Applying the max-sliced Wasserstein-2 distance yielded almost the same solution as the one with maximal energy for synthetic images (a linear correlation of 0.992). Notably, it is not straightforward to identify the other slice wherein natural images are over-represented, and synthetic images are under-represented, using the max-sliced Wasserstein-2 distance.

Figure 3: Max-slicing inception codes to illustrate the AutoGAN discrepancies. One-sided max-sliced Bures is used to identify two witness function (as linear subspaces of the inception codes) that differentiate the two samples. (. (Left: A,C) Images in subspace under-represented by AutoGAN. (Right: B,D) Images in subspace over-represented by AutoGAN. Top: A,B) CIFAR10 test images. (Bottom: C,D) Synthetic images from AutoGAN. (C) Realism scores (median and range): 0.92 (0.84–1.03). (D) Realism scores (median and range): 0.68 (0.62–0.73)

3.2 Covariate Shift Detection with Kernel-Based Divergences

We test whether the proposed kernel-based divergences are able to identify the particular instances implicated in a covariate shift between two samples. In this context, the witness function makes it straightforward to identify the specific instances in either sample that are associated with the imbalance. Those with extreme witness function evaluations are associated to under- or over-represented cases. A baseline approach[4] for this task is to train a logistic regression model with kernel basis functions to distinguish the two samples, and then use the probability estimates of the instances as the witness function evaluations $\omega(x) = \Pr(H_0|X=x) = 1-\Pr(H_1|X=x)$, where $H_0 : X \sim \hat{\nu}$ and $H_1 : X \sim \hat{\mu}$. We also compare with kernel Fischer discriminant analysis (KFDA) (Mika et al., 1999), and the linear cases of max-sliced Wasserstein-2 distance, its first moment approximation, max-sliced Bures, and logistic regression. For all kernel methods, a Gaussian kernel κ_σ is used with the parameter σ set as the median Euclidean distance in the pooled sample.

Using the MNIST data set, we test three scenarios of covariate shift. For each, one sample has a mismatched probability for one class $l \in \{0, \dots, 9\}$ and the other sample has a balanced sample: (Scenario 1) $\hat{\mu}$ is balanced and $\hat{\nu}$ is missing l; (Scenario 2) $\hat{\mu}$ is imbalanced with l only appearing in 2% of the cases, compared to $10.\bar{8}$% for the other classes and $\hat{\nu}$ is balanced; (Scenario 3) $\hat{\mu}$ is balanced and $\hat{\nu}$ consists of only images from l. In each case, $\hat{\mu}$ is a sample of 500 images from the training set and $\hat{\nu}$ is a sample of 500 images from the test set. A threshold-free way to assess these sorts of covariate shift are using the area-under-the-curve (AUC) of the receiver operator curve (ROC), where positive instances correspond to class l. For some methods, the witness function (or its magnitude) may be ambiguous in sign, i.e., the values may be high (or large) for either the under- or over-sampled instances. The ordering with the highest AUC on each run is chosen (see Table 2).

3.3 Covariate Shift Correction by Reweighting

We consider the task of re-weighting the instances in one sample to minimize the divergence. In the linear case for max-sliced Bures distance, this optimization can be expressed as $\min_{\boldsymbol{\nu} \in \mathbb{R}^n_{\geq 0} : \sum_i \nu_i = 1} J(\boldsymbol{\nu})$, where $J(\boldsymbol{\nu}) \propto \textit{max-}D_B^{\mathbb{R}^d}(\hat{\mu}, \hat{\nu})$. As shown in appendix A.5, this is a convex minimization with a simplex constraint on $\boldsymbol{\nu}$. We apply the Frank-Wolfe (conditional gradient)

[4] As additional baselines we also tested three methods for importance reweighting and density ratio estimation: kernel mean matching (KMM) (Huang et al., 2007), least-squares importance estimation (uLSIF) (Kanamori et al., 2009), and relative density-ratio estimation (RuLSIF) (Yamada et al., 2011) but all methods were outperformed by logistic regression with kernel bases.

Table 2: Unsupervised covariate shift outlier detection on MNIST. The goal is to identify instances associated with an over- or underrepresented class $l \in \{0, \ldots, 9\}$. Values are AUC where positives are instances from class l. We report the mean and standard deviation and the number of times each method has the highest AUC (including ties) across 100 trials (10 for each case $l \in \{0, \ldots 9\}$).

	(Scenario 1)	(Scenario 2)	(Scenario 3)	(1)	(2)	(3)
logistic regression-linear	0.60 (0.05)	0.60 (0.08)	0.91 (0.06)	0	1	0
Max-Sliced Bures	0.89 (0.12)	0.86 (0.13)	0.95 (0.03)	2	3	3
Max-Sliced W2 (approx.)	0.86 (0.10)	0.84 (0.13)	0.96 (0.02)	1	2	0
Max-Sliced W2	0.87 (0.14)	0.85 (0.14)	0.96 (0.02)	2	12	0
logistic regression-kernel	0.90 (0.07)	0.86 (0.10)	0.99 (0.01)	10	15	93
KFDA	0.58 (0.03)	0.61 (0.12)	0.91 (0.03)	0	2	1
MMD	0.87 (0.10)	0.85 (0.12)	0.97 (0.02)	1	3	0
Max-Sliced Kernel TV	0.85 (0.10)	0.83 (0.14)	0.96 (0.03)	2	0	1
Max-Sliced Kernel Bures	0.92 (0.11)	0.88 (0.14)	0.97 (0.02)	21	32	2
Max-Sliced Kernel W2	0.92 (0.11)	0.88 (0.13)	0.97 (0.02)	61	35	0

algorithm (Jaggi, 2013) to iteratively adjust the weight of one instance at each iteration. Figure 4 shows an example of the highest weighted instances, and compares re-weightings that minimize the Wasserstein-2 distance and the max-sliced Wasserstein-2 distance (using 10 mini-batch approximations of size $n = m =$100 at each Frank-Wolfe iteration). Figure 5 shows results of re-weighting a uniform distribution to match '25 Gaussians' using random Fourier bases (Rahimi & Recht, 2008).

Figure 4: Synthetic images with the largest values of weights for 50,000 synthetic images when the weights of $\hat{\nu}$ are optimized to match $\hat{\mu}$, which is the 'birds' subset of CIFAR10 training images.

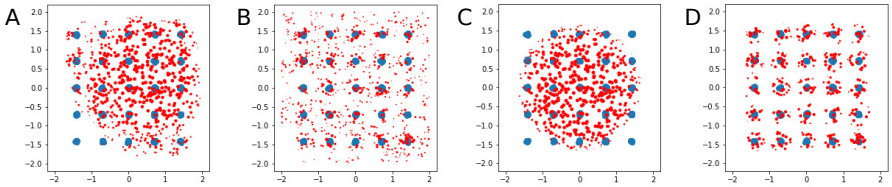

Figure 5: Re-weighting a uniform distribution to match a mixture of 25 Gaussians by minimizing the (A,B) max-sliced Wasserstein-2 distance and (C,D) max-sliced Bures distance with a (A,C) linear kernel or (B,D) non-linear random Fourier bases with d=2,000 with σ=0.2. Example follows Kolouri et al. (2019) and uses Adam (Kingma & Ba, 2015) with defaults besides a learning rate of 10^{-4}.

4 Conclusion

We have proposed the max-sliced Bures distance, a lower-bound on the max-sliced Wasserstein-2 distance, which is computationally simpler and performs similarly on tasks such as covariate shift detection. We show increased performance with kernel-based witness functions for covariate shift, and also highlight the linear case when the feature space is the internal representation of a pre-trained network. For linear mappings, the max-sliced Bures distance outperforms first-order approximations to the max-sliced Wasserstein-2 distance, and is much faster than the saddle-point optimization to achieve comparable performance. Additionally, the one-sided max-sliced Bures divergences enable direct interpretation of under- and over-representation between two samples, which can be used to identify systematic discrepancies such as mode dropping.

REFERENCES

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.

Theodore W Anderson. On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, pp. 1148–1159, 1962.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 214–223, 2017.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, pp. 337–404, 1950.

Thomas R. Bromley, Marco Cianciaruso, Rosario Lo Franco, and Gerardo Adesso. Unifying approach to the quantification of bipartite correlations by Bures distance. *Journal of Physics A: Mathematical and Theoretical*, 47(40):405302, 2014.

Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite W*-algebras. *Transactions of the American Mathematical Society*, pp. 199–212, 1969.

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*, pp. 3483–3491, 2018.

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*, pp. 10648–10656, 2019.

DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

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

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

Kenji Fukumizu, Arthur Gretton, Gert R Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, pp. 1750–1758, 2009.

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

Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. AutoGAN: Neural architecture search for generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems 19*, pp. 513–520. MIT Press, Cambridge, MA, 2007.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

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*, pp. 6626–6637, 2017.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pp. 601–608, 2007.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435, 2013.

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.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

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*, pp. 3427–3436, 2018.

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

Vladimir Koltchinskii and Dong Xia. Optimal estimation of low rank density matrices. *The Journal of Machine Learning Research*, 16:1757–1792, 2015.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pp. 3927–3936, 2019.

Gert R. Lanckriet and Bharath K. Sriperumbudur. On the convergence of the concave-convex procedure. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 22*, pp. 1759–1767. Curran Associates, Inc., 2009.

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*, pp. 10285–10295, 2019.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.

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

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pp. 41–48. IEEE, 1999.

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

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

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

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*, pp. 1089–1096, 2008.

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.

Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

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 (eds.), *Advances in Neural Information Processing Systems 29*, pp. 271–279. Curran Associates, Inc., 2016.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2008.

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.

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*, pp. 3723–3732, 2018.

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*, pp. 2234–2242, 2016.

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

Friedrich Schmid and Mark Trede. A distribution free test for the two sample problem for general alternatives. *Computational Statistics & Data Analysis*, 20(4):409–419, 1995.

Mark Schmidt. minFunc, 2012. Software available at `http://www.di.ens.fr/~mschmidt/Software/minFunc.html`.

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.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *21st Annual Conference on Learning Theory (COLT 2008)*, pp. 111–122. Omnipress, 2008.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

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*, pp. 2818–2826, 2016.

Pham Dinh Tao and Le Thi Hoai An. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Armin Uhlmann. The transition probability in the state space of a*-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976.

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

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*, pp. 3713–3722, 2019.

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*, pp. 594–602, 2011.

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.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4): 915–936, 2003.

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

The appendix details the derivation of the proposed divergences, formal results relating them to existing divergences, and additional experimental results.

A.1 Derivation of the max-sliced kernel divergences

Let d denote either the TV distance ($d = d_{TV}$) or the squared Bures distance ($d = d_B^2$), defined in equation 2 and equation 3, respectively. In these cases, the expression of the max-sliced kernel divergence is

$$\sup_{\Omega \in \mathcal{U}_1} d(\underbrace{\Omega \rho_X \Omega}_{\langle \Omega, \rho_X \rangle \Omega}, \underbrace{\Omega \rho_Y \Omega}_{\langle \Omega, \rho_Y \rangle \Omega}) = \sup_{\Omega \in \mathcal{U}_1} \delta(\langle \Omega, \rho_X \rangle, \langle \Omega, \rho_Y \rangle) = \sup_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \delta(\mathbb{E}\left[\omega^2(X)\right], \mathbb{E}\left[\omega^2(Y)\right]),$$

where $\delta(p, q) = d(p\Omega, q\Omega)$ denotes the distance between scaled versions of Ω, for $p = \langle \Omega, \rho_X \rangle \in \mathbb{R}_{\geq 0}$ and $q = \langle \Omega, \rho_Y \rangle \in \mathbb{R}_{\geq 0}$. The equalities follow from the facts that $\Omega \rho_X \Omega = (\omega \otimes \omega)\, \rho_X\, (\omega \otimes \omega) = \langle \Omega, \rho_X \rangle \Omega$, and $\langle \Omega, \rho_X \rangle = \mathbb{E}\left[\langle \phi(X), \omega \rangle \langle \phi(X), \omega \rangle\right] = \mathbb{E}\left[\omega^2(X)\right]$ and likewise for ρ_Y. The TV distance and squared Bures distance yield expressions for $\delta(p, q)$ that match the form of the underlying TV and Hellinger divergences: $d_{TV}(p\Omega, q\Omega)$ yields $\delta(p, q) = \frac{1}{2}|p - q|$, and $d_B^2(p\Omega, q\Omega)$ yields $\delta(p, q) = (\sqrt{p} - \sqrt{q})^2$. The different forms of the divergences and distances are listed in Table 3.

Slicing is natural for the operator based measures, as it is inherent in defining the operator based distances such that they coincide with the corresponding divergences between discrete probability laws (Fuchs & Van De Graaf, 1999). The scalar discrepancy measure $\delta(\cdot, \cdot)$ can be accumulated across a *complete* set slices to obtain the original distances. Consider a set (or countably infinite sequence) of orthogonal trace-norm operators $\mathcal{O} = \{\Omega_1 = \omega_1 \otimes \omega_1, \Omega_2 = \omega_2 \otimes \omega_2, \ldots, \Omega_k = \omega_k \otimes \omega_k\}$. Since these operators are orthogonal and trace-norm $\|\sum_{i=1}^k \Omega_i\|_\infty = 1$,

$$d(\rho_X, \rho_Y) \geq \sum_{i=1}^k d(\langle \Omega_i, \rho_X \rangle \Omega_i, \langle \Omega_i, \rho_Y \rangle \Omega_i) = \sum_{i=1}^k \delta(\langle \Omega_i, \rho_X \rangle, \langle \Omega_i, \rho_Y \rangle) = \sum_{i=1}^k \delta(p_i(\mathcal{O}), q_i(\mathcal{O})),$$

where $p_i(\mathcal{O}) = \langle \Omega_i, \rho_X \rangle$ and $q_i(\mathcal{O}) = \langle \Omega_i, \rho_Y \rangle$. To obtain the equality, one must optimize over all possible sets of operators $\mathcal{O}$.

Table 3: Relationship between divergence D, operator dissimilarity d, and scalar discrepancy δ for the TV and squared Bures/Hellinger divergences (squared Bures is twice the squared Hellinger).

	$D(\mu,\nu)$	$d(\rho_X,\rho_Y)$	$\delta(p,q)$
TV	$\frac{1}{2}\int_{\mathcal{X}}\lvert d\mu(x)-d\nu(x)\rvert$	$\frac{1}{2}\lVert\rho_X-\rho_Y\rVert_1$	$\frac{1}{2}\lvert p-q\rvert$
(Bures)^2	$\int_{\mathcal{X}}\left(\sqrt{d\mu(x)}-\sqrt{d\nu(x)}\right)^2$	$\lVert\rho_X\rVert_1+\lVert\rho_Y\rVert_1-2\left\lVert\rho_X^{\frac{1}{2}}\rho_Y^{\frac{1}{2}}\right\rVert_1$	$\lvert\sqrt{p}-\sqrt{q}\rvert^2$

A.1.1 MAX SLICED TOTAL VARIATION DISTANCE

The sliced kernel total variation distance is

$$d_{TV}(\Omega\rho_X\Omega,\Omega\rho_Y\Omega)=\frac{1}{2}\|\Omega(\rho_X-\rho_Y)\Omega\|_1=\frac{1}{2}\|\langle\Omega,\rho_X-\rho_Y\rangle\Omega\|_1=\frac{1}{2}|\langle\Omega,\rho_X-\rho_Y\rangle|\|\Omega\|_1.$$

Maximizing over slices yields

$$\frac{1}{2}\sup_{\Omega\in\mathcal{U}_1}|\langle\Omega,\rho_X-\rho_Y\rangle|\|\Omega\|_1=\frac{1}{2}\sup_{\Omega\in\mathcal{U}_1}|\langle\Omega,\rho_X-\rho_Y\rangle|=\frac{1}{2}\sup_{\omega:\|\omega\|_2\leq 1}|\|\omega\|_\mu^2-\|\omega\|_\nu^2|, \tag{12}$$

where the first equality follows from the fact that distance is maximized when $\|\Omega\|_1=1$. This yields the expression

$$\textit{max-}D_{TV}^{\mathcal{H}}(\mu,\nu)\triangleq\frac{1}{2}\sup_{\omega:\|\omega\|_2\leq 1}|\|\omega\|_\mu^2-\|\omega\|_\nu^2|. \tag{13}$$

Notably, the penultimate expression in equation 12 can be related to the operator norm,

$$\sup_{\Omega\in\mathcal{U}_1}|\langle\Omega,\rho_X-\rho_Y\rangle|\leq\sup_{\Omega\in\{O\in\mathcal{H}\times\mathcal{H}:\|O\|_1\leq 1\}}\langle\Omega,\rho_X-\rho_Y\rangle=\|\rho_X-\rho_Y\|_\infty, \tag{14}$$

due to the dual norm definition. Thus, *max-*$D_{TV}^{\mathcal{H}}(\mu,\nu)\leq\frac{1}{2}\|\rho_X-\rho_Y\|_\infty$. In the linear case, this is an equality as the optimization corresponds to finding the largest magnitude eigenvalue.

A.1.2 MAX SLICED BURES DISTANCE

The sliced version of the Bures distance is

$$d_B(\Omega\rho_X\Omega,\Omega\rho_Y\Omega)=\sqrt{\|\Omega\rho_X\Omega\|_1+\|\Omega\rho_Y\Omega\|_1-2\|(\Omega\rho_X\Omega)^{\frac{1}{2}}(\Omega\rho_Y\Omega)^{\frac{1}{2}}\|_1},$$

which can be simplified since $\|\Omega\rho_X\Omega\|_1=\mathrm{tr}((\omega\otimes\omega)\rho_X\omega\otimes\omega)=\langle\omega,\rho_X\omega\rangle\|\omega\|_2^2=\|\omega\|_\mu^2\|\omega\|_2^2$, and $\|(\Omega\rho_X\Omega)^{\frac{1}{2}}(\Omega\rho_Y\Omega)^{\frac{1}{2}}\|_1=\sqrt{\omega_\mu^2\omega_\nu^2}\|\Omega\|_1=\|\omega\|_\mu\|\omega\|_\nu\|\omega\|_2^2$. Using these expressions, the max-sliced Bures distance is

$$\sup_{\Omega\in\mathcal{U}_1}d_B(\Omega\rho_X\Omega,\Omega\rho_Y\Omega)=\sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1}\|\omega\|_2\sqrt{\|\omega\|_\mu^2+\|\omega\|_\nu^2-2\|\omega\|_\mu\|\omega\|_\nu}.$$

The expression is monotonic with the norm of ω yielding

$$\begin{aligned}\textit{max-}D_B^{\mathcal{H}}(\mu,\nu)&\triangleq\sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1}\sqrt{(\|\omega\|_\mu-\|\omega\|_\nu)^2}=\sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1}\left|\|\omega\|_\mu-\|\omega\|_\nu\right| \\ &=\sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1}\left|\sqrt{\mathbb{E}_{X\sim\mu}[\omega^2(X)]}-\sqrt{\mathbb{E}_{Y\sim\nu}[\omega^2(Y)]}\right|.\end{aligned} \tag{15}$$

A.1.3 MAX SLICED GAUSS-WASSERSTEIN DISTANCES

Two max-sliced versions of the Gauss-Wasserstein distance in the RKHS are

$$\begin{aligned}
\textit{max}_L\text{-}D_{GW}^{\mathcal{H}}(\mu,\nu) &= \sup_{\Omega\in\mathcal{U}_1} \sqrt{\|\Omega(m_X - m_Y)\|_2^2 + d_B^2(\Omega\Sigma_X\Omega, \Omega\Sigma_Y\Omega)} \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \|\omega\|_2 \sqrt{\langle m_X - m_Y, \omega\rangle^2 + \left(\sqrt{\langle\omega, \Sigma_X\omega\rangle} - \sqrt{\langle\omega, \Sigma_Y\omega\rangle}\right)^2} \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \sqrt{\langle m_X - m_Y, \omega\rangle^2 + \left(\sqrt{\langle\omega, \Sigma_X\omega\rangle} - \sqrt{\langle\omega, \Sigma_Y\omega\rangle}\right)^2} \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \sqrt{(\mathbb{E}_{X\sim\mu}[\omega(X)] - \mathbb{E}_{Y\sim\nu}[\omega(Y)])^2 + \left(\sigma_{\omega(X)} - \sigma_{\omega(Y)}\right)^2}, \quad (16)
\end{aligned}$$

where $\sigma_{\omega(X)} = \sqrt{\mathbb{E}[(\omega(X) - \langle m_X, \omega\rangle)^2]}$, is the standard deviation of $\omega(X)$, and $\sigma_{\omega(Y)} = \sqrt{\mathbb{E}[(\omega(Y) - \langle m_Y, \omega\rangle)^2]}$, and

$$\begin{aligned}
\textit{max}_U\text{-}D_{GW}^{\mathcal{H}}(\mu,\nu) &= \left(\sup_{\Omega\in\mathcal{U}_1} \|\Omega(m_X - m_Y)\|_2^2 + \sup_{\Omega\in\mathcal{U}_1} d_B^2(\Omega\Sigma_X\Omega, \Omega\Sigma_Y\Omega)\right)^{\frac{1}{2}} \\
&= \sqrt{\textit{MMD}^2(\mu,\nu) + \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \left(\sigma_{\omega(X)} - \sigma_{\omega(Y)}\right)^2}, \quad (17)
\end{aligned}$$

which is an upper bound on MMD. Notably, $\textit{max}_L\text{-}D_{GW}^{\mathcal{H}}(\mu,\nu)$ is the supremum of the Fréchet distance over all witness functions.

A.1.4 MAX-SLICED KERNEL WASSERSTEIN

A sliced version of the kernel Wasserstein distance between μ and ν relies on the the sliced distance

$$\begin{aligned}
d_{\kappa,\omega}(X,Y) &= \|(\omega\otimes\omega)(\phi(X) - \phi(Y))\|_2 = \|\langle\omega, \phi(X) - \phi(Y)\rangle\omega\|_2 \\
&= |\langle\omega, \phi(X) - \phi(Y)\rangle|\|\omega\|_2 = |\langle\omega, \phi(X)\rangle - \langle\omega, \phi(Y)\rangle|\|\omega\|_2 = |\omega(X) - \omega(Y)|\|\omega\|_2.
\end{aligned}$$

This distance is monotonic with the norm of ω and convex with respect to ω. The max-sliced kernel Wasserstein-p distance, $p \geq 1$, is

$$\begin{aligned}
\textit{max-}W_p^{\mathcal{H}}(\mu,\nu) &= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \inf_{\gamma\in\Gamma(\mu,\nu)} \left[\mathop{\mathbb{E}}_{(X,Y)\sim\gamma} (d_{\kappa,\omega}(X,Y))^p\right]^{\frac{1}{p}} \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \inf_{\gamma\in\Gamma(\mu,\nu)} \left[\mathop{\mathbb{E}}_{(X,Y)\sim\gamma} |\omega(X) - \omega(Y)|^p\right]^{\frac{1}{p}}, \quad (18)
\end{aligned}$$

which is a one-dimensional optimal transport problem. Let $\omega_{\#}\mu$ and $\omega_{\#}\nu$ denote the pushforward measures, then the divergence can be written as

$$\textit{max-}W_p^{\mathcal{H}}(\mu,\nu) = \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \inf_{\pi\in\Pi(\omega_{\#}\mu,\omega_{\#}\nu)} \left[\mathop{\mathbb{E}}_{(S,T)\sim\pi} |S - T|^p\right]^{\frac{1}{p}}, \quad (19)$$

where $\Pi(\omega_{\#}\mu, \omega_{\#}\nu)$ is the set of all joint distributions coupling the pushforward measures.

Assuming the measures are absolutely continuous and adopting the notation from Santambrogio (2015), and let $F_{\omega,\mu}(w) = \int_{-\infty}^{w} d\omega_{\#}\mu = \omega_{\#}\mu((-\infty, w])$ and $F_{\omega,\nu}(w) = \int_{-\infty}^{w} d\omega_{\#}\nu = \omega_{\#}\nu((-\infty, w])$ denote the cumulative distribution functions of the pushforward measures with pseudo-inverses $F_{\omega,\mu}^{-1}(q) = \inf\{w \in \mathbb{R} : F_{\omega,\mu}(w) \geq q\}$ and $F_{\omega,\nu}^{-1}(q) = \inf\{w \in \mathbb{R} : F_{\omega,\nu}(w) \geq q\}$. As shown in Lemma 2.8 (Santambrogio, 2015), then the optimal transport plan π^* has cumulative distribution $G_\omega(w_X, w_Y) = \min\{F_{\omega,\mu}(w_X), F_{\omega,\nu}(w_Y)\}$ and the divergence is

$$\textit{max-}W_p^{\mathcal{H}}(\mu,\nu) = \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \left[\int_0^1 \left|F_{\omega,\mu}^{-1}(q) - F_{\omega,\nu}^{-1}(q)\right|^p dq\right]^{\frac{1}{p}}, \quad (20)$$

and for the case $p = 1$ (Santambrogio, 2015, Proposition 2.17),

$$\textit{max-}W_1^{\mathcal{H}}(\mu,\nu) \triangleq \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \int_{\mathbb{R}} |F_{\omega,\mu}(w) - F_{\omega,\nu}(w)|\, dw. \tag{21}$$

The objective in the last quantity is an L_1 norm version of the L_2 Cramér–von Mises criterion (Schmid & Trede, 1995; Anderson, 1962). The max-sliced kernel Wasserstein-1 distance is equivalent to a max-sliced L_1 norm version of the Cramér–von Mises criterion, where the slicing corresponds to a function in the RKHS that witnesses the largest discrepancies between the measures. The choice of $p = 2$ simplifies, yielding

$$\begin{aligned}
\textit{max-}W_2^{\mathcal{H}}(\mu,\nu) &= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\mathop{\mathbb{E}}_{(X,Y)\sim\gamma} \left[\omega^2(X) + \omega^2(Y) - 2\omega(X)\omega(Y)\right] \right)^{\frac{1}{2}} \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \left(\|\omega\|_\mu^2 + \|\omega\|_\nu^2 - \sup_{\gamma\in\Gamma(\mu,\nu)} \mathop{\mathbb{E}}_{(X,Y)\sim\gamma} \left[2\omega(X)\omega(Y)\right] \right)^{\frac{1}{2}} && (22) \\
&= \sup_{\omega\in\mathcal{H}:\|\omega\|_2\leq 1} \left(\|\omega\|_\mu^2 + \|\omega\|_\nu^2 - 2\int_0^1 F^{-1}_{\omega,\mu}(q) F^{-1}_{\omega,\nu}(q) dq \right)^{\frac{1}{2}}. && (23)
\end{aligned}$$

A.2 Relationship between the kernel-based max-sliced divergences

The Bures distance can be used to lower bound the TV distance, $d_B^2(\rho_X,\rho_Y) \leq \|\rho_X - \rho_Y\|_1 = 2d_{TV}(\rho_X,\rho_Y)$ (Fuchs & Van De Graaf, 1999). This inequality stems from the inequality between the squared Hellinger and total variation distances for discrete probability laws. $\sum_i \|\sqrt{p_i} - \sqrt{q_i}\|_2^2 \leq \sum_i |p_i - q_i|$, where the inequality holds for each summand, $|\sqrt{p_i} - \sqrt{q_i}|^2 \leq |\sqrt{p_i} - \sqrt{q_i}|(\sqrt{p_i} + \sqrt{q_i}) = |p_i - q_i|$. This inequality holds for the max-sliced measures, as stated in the following theorem.

Theorem 1. *The square of max-sliced Bures distance in the RKHS induced by the kernel κ is less than or equal to the twice the max-sliced TV distance* $\left(\textit{max-}D_B^{\mathcal{H}}(\mu,\nu)\right)^2 \leq 2\left(\textit{max-}D_{TV}^{\mathcal{H}}(\mu,\nu)\right)$.

Proof of Theorem 1. $\left(\|\omega\|_\mu - \|\omega\|_\nu\right)^2 \leq \left|\|\omega\|_\mu^2 - \|\omega\|_\nu^2\right|$, where the inequality is a consequence of the inequality between the arithmetic and geometric means, let $a = \|\omega\|_\mu^2$ and $b = \|\omega\|_\nu^2$, then $(\sqrt{a} - \sqrt{b})^2 \leq |\sqrt{a} - \sqrt{b}|(\sqrt{a} + \sqrt{b}) = |a - b|$. Taking the supremum over $\omega \in \mathcal{H}$ with the constraint $\|\omega\|_2 \leq 1$ yields the desired the result $\left(\textit{max-}D_B^{\mathcal{H}}(\mu,\nu)\right)^2 \leq 2\left(\textit{max-}D_{TV}^{\mathcal{H}}(\mu,\nu)\right)$.

□

We now consider the relationship between the Gauss-Wasserstein or Fréchet distance, which combine the distances between the first and second-order moments, with the max-sliced Bures, which is applied directly to the uncentered covariance matrices. For this we need the following lemma.

Lemma 2. *For two vectors in $\mathbb{R}^2$, the difference between their Euclidean norms is less than or equal to the Euclidean norm of their differences. For $a, b, c, d \in \mathbb{R}$, $\left|\sqrt{a^2+b^2} - \sqrt{c^2+d^2}\right| \leq \sqrt{(a-c)^2 + (b-d)^2}$.*

Proof. Let $e = \left(\sqrt{a^2+b^2} - \sqrt{c^2+d^2}\right)^2$ and $f = (a-c)^2 + (b-d)^2$.

$$\begin{aligned}
e &= a^2 + b^2 + c^2 + d^2 - 2\sqrt{(a^2+b^2)(c^2+d^2)} \\
&= a^2 + b^2 + c^2 + d^2 - 2\sqrt{a^2c^2 + b^2d^2 + b^2c^2 + a^2d^2} \\
&\leq a^2 + b^2 + c^2 + d^2 - 2\sqrt{a^2c^2 + b^2d^2 + 2\sqrt{a^2b^2c^2d^2}}, \text{ by the arithmetic and geometric mean inequality} \\
&= a^2 + b^2 + c^2 + d^2 - 2\sqrt{(\sqrt{a^2c^2} + \sqrt{b^2d^2})^2} \\
&= a^2 + b^2 + c^2 + d^2 - 2(\sqrt{a^2c^2} + \sqrt{b^2d^2}) \\
&\leq a^2 + b^2 + c^2 + d^2 - 2(ac + bd) = (a-c)^2 + (b-d)^2 = f.
\end{aligned}$$

Taking the square root of each side yields the inequality $\sqrt{e} \leq \sqrt{f}$. □

Theorem 3. *The kernel max-sliced Bures distance in the RKHS $\mathcal{H}$ is a lower bound on the kernel max-sliced Gauss-Wasserstein distance, max-$D_B^{\mathcal{H}}(\mu,\nu) \leq$ max$_L$-$D_{GW}^{\mathcal{H}}(\mu,\nu) \leq$ max$_U$-$D_{GW}^{\mathcal{H}}(\mu,\nu)$.*

Proof. To relate the sliced Bures and Gauss-Wasserstein distances, we note that $\|\omega\|_\mu^2 = \langle \rho_X, \omega \otimes \omega \rangle = \langle \Sigma_X + m_X \otimes m_X, \omega \otimes \omega \rangle = \sigma_{\omega(X)}^2 + \langle m_X, \omega \rangle^2$ and likewise for $\|\omega\|_\nu^2$. Then

$$\begin{aligned} \left| \sqrt{\|\omega\|_\mu^2} - \sqrt{\|\omega\|_\nu^2} \right| &= \left| \sqrt{\langle m_X, \omega \rangle^2 + \sigma_{\omega(X)}^2} - \sqrt{\langle m_Y, \omega \rangle^2 + \sigma_{\omega(Y)}^2} \right| \\ &\leq \sqrt{\langle m_X - m_Y, \omega \rangle^2 + (\sigma_{\omega(X)} - \sigma_{\omega(Y)})^2}, \end{aligned} \tag{24}$$

where the inequality relies on Lemma 2, with $a = \langle m_X, \omega \rangle$, $b = \sigma_{\omega(X)}$, $c = \langle m_Y, \omega \rangle$, and $d = \sigma_{\omega(Y)}$. Taking supremum over slices yields the desired inequality. □

This shows that the max-sliced kernel Bures distance is a lower-bound on the max-sliced kernel Wasserstein-2 distance, since the latter is lower bounded by the max-sliced kernel Gauss-Wasserstein distance.

Theorem 4. *The kernel max-sliced Bures distance in the RKHS $\mathcal{H}$ is a lower bound on the kernel max-sliced Wasserstein 2 distance, max-$D_B^{\mathcal{H}}(\mu,\nu) \leq$ max-$W_2^{\mathcal{H}}(\mu,\nu)$.*

Proof. Let $\omega(X) = \langle m_X, \omega \rangle + \tilde{\omega}(X)$ and $\omega(Y) = \langle m_Y, \omega \rangle + \tilde{\omega}(Y)$, where $\tilde{\omega}(X) = \langle \phi(X) - m_X, \omega \rangle$ and $\tilde{\omega}(Y) = \langle \phi(Y) - m_Y, \omega \rangle$ are zero mean, and $\mathbb{E}\left[\tilde{\omega}^2(X)\right] = \sigma_{\omega(X)}^2$ and $\mathbb{E}\left[\tilde{\omega}^2(Y)\right] = \sigma_{\omega(Y)}^2$. The squared Fréchet distance between random variables $\omega(X)$ and $\omega(Y)$ is

$$\begin{aligned} &(\mathbb{E}\left[\omega(X)\right] - \mathbb{E}\left[\omega(Y)\right])^2 + (\sigma_{\omega(X)} - \sigma_{\omega(Y)})^2 = \langle m_X - m_Y, \omega \rangle^2 + (\sigma_{\omega(X)} - \sigma_{\omega(Y)})^2 \\ &= \langle m_X, \omega \rangle^2 + \sigma_{\omega(X)}^2 + \langle m_Y, \omega \rangle^2 + \sigma_{\omega(Y)}^2 - 2\left(\langle m_X, \omega \rangle \langle m_Y, \omega \rangle + \sigma_{\omega(X)} \sigma_{\omega(Y)} \right) \\ &= \mathbb{E}\left[\omega^2(X)\right] + \mathbb{E}\left[\omega^2(Y)\right] - 2\left(\mathbb{E}\left[\omega(X)\right] \mathbb{E}\left[\omega(Y)\right] + \sqrt{\mathbb{E}\left[\tilde{\omega}^2(X)\right] \mathbb{E}\left[\tilde{\omega}^2(Y)\right]} \right). \end{aligned}$$

By Hölder's inequality, $\mathbb{E}\left[\tilde{\omega}(X)\tilde{\omega}(Y)\right] \leq \sigma_{\omega(X)}\sigma_{\omega(Y)} = \sqrt{\mathbb{E}\left[\tilde{\omega}^2(X)\right] \mathbb{E}\left[\tilde{\omega}^2(Y)\right]}$. Consequently,

$$\mathbb{E}\left[\omega^2(X)\right] + \mathbb{E}\left[\omega^2(Y)\right] - 2\left(\mathbb{E}\left[\omega(X)\right] \mathbb{E}\left[\omega(Y)\right] + \sqrt{\mathbb{E}\left[\tilde{\omega}^2(X)\right] \mathbb{E}\left[\tilde{\omega}^2(Y)\right]} \right) \tag{25}$$

$$\leq \mathbb{E}\left[\omega^2(X)\right] + \mathbb{E}\left[\omega^2(Y)\right] - 2\left(\mathbb{E}\left[\omega(X)\right] \mathbb{E}\left[\omega(Y)\right] + \mathbb{E}\left[\tilde{\omega}(X)\tilde{\omega}(Y)\right] \right)$$

$$= \mathbb{E}\left[\omega^2(X)\right] + \mathbb{E}\left[\omega^2(Y)\right] - 2\,\mathbb{E}\left[\omega(X)\omega(Y)\right] = \mathbb{E}\left[(\omega(X) - \omega(Y))^2\right]. \tag{26}$$

Taking the infimum over all possible joint distributions $(X, Y) \sim \gamma$ that are within the coupling distribution $\gamma \in \Gamma$, yields the sliced Wasserstein-2 distance on the right hand side. Maximizing over slices, yields max$_U$-$D_{GW}^{\mathcal{H}}(\mu,\nu) \leq$ max-$W_2^{\mathcal{H}}(\mu,\nu)$ and combining with Theorem 3 yields the desired result. □

In the linear case with $\omega(X) = \langle \mathbf{w}, X \rangle$, $\omega(Y) = \langle \mathbf{w}, Y \rangle$ and $\mathbf{w}, X, Y \in \mathbb{R}^d$, this shows that the max-sliced Bures distance is a lower-bound on the max-sliced Fréchet distance, which is a lower-bound on the max-sliced Wasserstein-2 distance.

A.3 Relationship to Other Divergences

Finally, we note that the terms in the max-sliced Gauss-Wasserstein divergences are related to kernel Fischer discriminant analysis (KFDA) (Mika et al., 1999). KFDA objective is based on the ratio of the difference in means to the pooled variances:

$$D_{FDA}^{\mathcal{H}}(\mu,\nu) = \sup_\omega \frac{\langle \omega, m_X - m_Y \rangle^2}{\sigma_{\omega(X)}^2 + \sigma_{\omega(Y)}^2} = \sup_\omega \frac{\langle \omega, m_X - m_Y \rangle^2}{\langle \omega, (\Sigma_X + \Sigma_Y)\omega \rangle}. \tag{27}$$

KFDA seeks a witness function which has widely separated means for the two measures, and minimal variance.

A.4 OPTIMIZATIONS FOR THE MAX-SLICED KERNEL DIVERGENCES

Due to the representer theorem for the RKHS, the witness function $\omega \in \mathcal{H}$ is of the form $\omega = \sum_{i=1}^{n+m} \alpha_i \phi(z_i)$ where $z_k = \begin{cases} x_k, & 1 \leq k \leq m \\ y_{k-m}, & m < k \leq n+m \end{cases}$ and $\boldsymbol{\alpha} \in \mathbb{R}^{m+n}$. In this case,

$$\begin{aligned}
\|\omega\|_{\hat{\mu}}^2 &= \langle \omega \otimes \omega, (\phi \otimes \phi)_{\#}\hat{\mu})\rangle = \sum_{j=1}^{m} \mu_j \langle \omega \otimes \omega, \phi(x_j) \otimes \phi(x_j)\rangle = \sum_{j=1}^{m} \mu_j \langle \omega, \phi(x_j)\rangle^2 \\
&= \sum_{j=1}^{m} \mu_j \langle \sum_{i=1}^{n+m} \alpha_i \phi(z_i), \phi(x_j)\rangle^2 = \sum_{j=1}^{m} \mu_j \left(\sum_{i=1}^{n+m} \alpha_i \kappa(z_i, x_j) \right)^2 = \langle \boldsymbol{\mu}, (\mathbf{K}_{XZ}\boldsymbol{\alpha})^{\circ 2}\rangle \\
&= \boldsymbol{\alpha}^\top \mathbf{K}_{XZ}^\top \text{diag}(\boldsymbol{\mu}) \mathbf{K}_{XZ} \boldsymbol{\alpha} = \|\mathbf{D}_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2^2,
\end{aligned}$$

where $\kappa(z_i, z_j) = K_{ij}$, $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{XX}, \mathbf{K}_{XY} \\ \mathbf{K}_{YX}, \mathbf{K}_{YY} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{XZ} \\ \mathbf{K}_{YZ} \end{bmatrix}$, $(\cdot)^{\circ 2}$ denotes the elementwise squaring of a matrix/vector, and $\mathbf{D}_{\mathbf{v}} = \text{diag}(\mathbf{v})$ denotes a diagonal matrix whose diagonal entries are the vector $\mathbf{v}$. Similarly, $\|\omega\|_{\hat{\nu}}^2 = \langle \boldsymbol{\nu}, (\mathbf{K}_{YZ}\boldsymbol{\alpha})^{\circ 2}\rangle = \|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2^2$.

A.4.1 MAX-SLICED KERNEL TV DISTANCE

Using these expressions, the max-sliced kernel TV distance is

$$\begin{aligned}
\textit{max-}D_{TV}^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) &= \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \frac{1}{2} \left| \|\mathbf{D}_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2^2 - \|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2^2 \right| && (28) \\
&= \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \frac{1}{2} |\boldsymbol{\alpha}^\top (\mathbf{K}_{XZ}^\top \mathbf{D}_{\boldsymbol{\mu}} \mathbf{K}_{XZ} - \mathbf{K}_{YZ}^\top \mathbf{D}_{\boldsymbol{\nu}} \mathbf{K}_{YZ}) \boldsymbol{\alpha}|. && (29)
\end{aligned}$$

The solution to this problem is the eigenvector corresponding to the largest magnitude eigenvalue of the generalized eigenvalue problem $\mathbf{A}\mathbf{v} = \lambda \mathbf{K}\mathbf{v}$, where $\mathbf{A} = \mathbf{K}_{XZ}^\top \mathbf{D}_{\boldsymbol{\mu}} \mathbf{K}_{XZ} - \mathbf{K}_{YZ}^\top \mathbf{D}_{\boldsymbol{\nu}} \mathbf{K}_{YZ}$.

A.4.2 MAX-SLICED KERNEL BURES DISTANCE

The max-sliced kernel Bures distance is

$$\begin{aligned}
\textit{max-}D_{B}^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) &= \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \left| \|\mathbf{D}_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2 - \|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2 \right| && (30) \\
&= \max_{s \in \{-1,+1\}} \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} s\|\mathbf{D}_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2 - s\|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2 && (31)
\end{aligned}$$

The last expression shows that the max-sliced Bures distance can be expressed as bilevel optimization, where the inner optimization, which we refer to as one-sided max-sliced Bures divergence, is a difference of convex functions:

$$\textit{max-}D_{B}^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \max \left\{ \left[\min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} g(\boldsymbol{\alpha}) - h(\boldsymbol{\alpha}) \right], \left[\min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} h(\boldsymbol{\alpha}) - g(\boldsymbol{\alpha}) \right] \right\} \tag{32}$$

where

$$g(\boldsymbol{\alpha}) = \|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2, \tag{33}$$

$$h(\boldsymbol{\alpha}) = \|\mathbf{D}_{\boldsymbol{\mu}}^{\frac{1}{2}} \mathbf{K}_{XZ} \boldsymbol{\alpha}\|_2. \tag{34}$$

This optimization can be solved as a minimization of a difference of convex functions, i.e. DC programming (Tao & An, 1998; Yuille & Rangarajan, 2003; Lanckriet & Sriperumbudur, 2009). Without loss of generality, we will consider the case first case in the maximum ($s = 1$),

$$\min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} g(\boldsymbol{\alpha}) - h(\boldsymbol{\alpha}) = \max_{\boldsymbol{\beta}} g^*(\boldsymbol{\beta}) - h^*(\boldsymbol{\beta}). \tag{35}$$

where $g^*(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle - g(\boldsymbol{\alpha})$ and $h^*(\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle - h(\boldsymbol{\alpha})$ are the Fenchel conjugate functions. To solve this, we use the difference of convex algorithm (DCA) (Tao & An, 1998), which is an alternating algorithm:

Choose $\boldsymbol{\alpha}_{(0)} \in \{\boldsymbol{\alpha} : \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1\}$ (feasible primal initialization)
$k \leftarrow 0$
while not converged **do**
  $\boldsymbol{\beta}_{(k)} \in \partial h(\boldsymbol{\alpha}_{(k)})$, (dual variable step)
  choose $\boldsymbol{\alpha}_{(k+1)} \in \partial g^*(\boldsymbol{\beta}_{(k)})$, (primal variable step)
  k++
end while
return $\alpha_{(k)}$

For $\boldsymbol{\alpha}_{(k)} \in \{\boldsymbol{\alpha} | h(\boldsymbol{\alpha}) > 0\}$, the dual variable step is $\partial h(\boldsymbol{\alpha}_{(k)}) = \frac{1}{h(\boldsymbol{\alpha}_{(k)})} \mathbf{K}_{XZ}^\top \mathbf{D}_{\boldsymbol{\mu}} \mathbf{K}_{XZ} \boldsymbol{\alpha}_{(k)}$. If $\boldsymbol{\alpha}_{(k)}$ lies in the nullspace of $\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ}$ then $h(\boldsymbol{\alpha}_{(k)})$ is not differentiable; however, as the optimization seeks to maximize $h(\boldsymbol{\alpha})$ this is avoidable.

The primal variable step corresponds to $\partial g^*(\boldsymbol{\beta}_{(k)}) = \arg\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}_{(k)} \rangle - g(\boldsymbol{\alpha})$, which is a convex quadratically constrained quadratic program

$$\min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1} \quad \|\mathbf{D}_{\boldsymbol{\nu}}^{\frac{1}{2}} \mathbf{K}_{YZ} \boldsymbol{\alpha}\|_2 + \langle \boldsymbol{\alpha}, -\boldsymbol{\beta}_{(k)} \rangle, \tag{36}$$

which can be solved by a standard convex solver (it can be reformulated as a second-order cone problem). Alternatively, by using smoothing and a log-barrier method to enforce the constraint that $\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} \leq 1$, the problem can be solved as a series of unconstrained non-linear optimizations with first-order solvers employing quasi-Newton approaches. In particular, we implement the barrier method in MATLAB and use MINFUNC (Schmidt, 2012) for unconstrained optimization.

A.4.3 MAX-SLICED KERNEL WASSERSTEIN

The empirical version of max-sliced kernel Wasserstein divergence can be expressed in terms of either equation 18 or equation 20,

$$\textit{max-}W_p^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \max_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \min_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \left[\sum_{i=1,j=1}^{m,n} P_{i,j} \left| \omega(x_i) - \omega(y_j) \right|^p \right]^{\frac{1}{p}} \tag{37}$$

$$= \max_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \left[\int_0^1 |F_{\omega,\hat{\mu}}^{-1}(q) - F_{\omega,\hat{\nu}}^{-1}(q)|^p dq \right]^{\frac{1}{p}}, \tag{38}$$

where $\mathcal{P}_{\hat{\mu},\hat{\nu}} = \{\mathbf{P} \in [0,1]^{m \times n} | \mathbf{P}\mathbf{1}_n = \boldsymbol{\mu}, \mathbf{P}^\top \mathbf{1}_m = \boldsymbol{\nu}\}$ is a transportation polytope. The empirical distribution functions are $F_{\omega,\hat{\mu}}(w) = \sum_{i=1}^m \mu_i \mathbb{I}_{\omega(x_i) \leq w}$ and $F_{\omega,\hat{\nu}}(w) = \sum_{i=1}^n \nu_i \mathbb{I}_{\omega(y_i) \leq w}$ with inverses $F_{\omega,\hat{\mu}}^{-1}(q) = \min\{\omega(x_i), i \in \{1, \ldots, m\} : F_{\omega,\hat{\mu}}(\omega(x_i)) \geq q\}$ and $F_{\omega,\hat{\nu}}^{-1}(q) = \min\{\omega(y_i), i \in \{1, \ldots, n\} : F_{\omega,\hat{\nu}}(\omega(y_i)) \geq q\}$. For fixed ω, the optimal transport plan $\acute{\mathbf{P}}$ is based on the sorted values $\omega(x_{(1)}) \leq \omega(x_{(2)}) \leq \cdots \leq \omega(x_{(m)})$ and $\omega(y_{(1)}) \leq \omega(y_{(2)}) \leq \cdots \leq \omega(y_{(m)})$. Denote the sorted values as vectors $\boldsymbol{\omega}'_X$ and $\boldsymbol{\omega}'_Y$, with $[\boldsymbol{\omega}'_X]_i = \omega(x_{(i)})$ and $[\boldsymbol{\omega}'_Y]_i = \omega(y_{(i)})$, and denote by $\acute{\boldsymbol{\mu}}$ and $\acute{\boldsymbol{\nu}}$ the corresponding permutations of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$.

In the case of equal samples sizes of uniform measure, $m = n$ and $\boldsymbol{\mu} = \frac{1}{m}\mathbf{1}_m$ and $\boldsymbol{\nu} = \frac{1}{n}\mathbf{1}_n$, elements in $\mathcal{P}_{\hat{\mu},\hat{\nu}}$ are scaled elements in the Birkhoff polytope (the set of doubly stochastic matrices), and the solution to the linear program is a permutation and $\textit{max-}W_p^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \sup_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \left\{ \left[\sum_{i=1}^m \left| \omega(x_{(i)}) - \omega(y_{(i)}) \right|^p \right]^{\frac{1}{p}} = \|\boldsymbol{\omega}'_X - \boldsymbol{\omega}'_Y\|_p \right\}$, where $\|\cdot\|_p$ denotes the ℓ_p norm in m dimensions. As in the continuous case, the discrete transport plan between the sorted measures has the cumulative distribution $\acute{\mathbf{G}} \in [0,1]^{m \times n}$ with $\acute{G}_{i,j} = \min\{\sum_{k=1}^i \acute{\mu}_k, \sum_{k=1}^j \acute{\nu}_k\}$. The optimal transport plan between the sorted measures is given by taking the first difference over both rows and columns of $\acute{\mathbf{G}}$, $\acute{P}_{1,1} = \acute{G}_{1,1}$, $\acute{P}_{1,j} = \acute{G}_{1,j} - \acute{G}_{1,j-1}$, $\acute{P}_{i,1} = \acute{G}_{i,1} - \acute{G}_{i-1,1}$, and $\acute{P}_{i,j} = \acute{P}_{i,j} - \acute{P}_{i-1,j} - \acute{P}_{i,j-1}$. Using the sorted witness function evaluations the distance can be written as

$$\textit{max-}W_p^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \max_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \left[\sum_{i=1,j=1}^{m,n} \acute{P}_{i,j} \left| \omega(x_{(i)}) - \omega(y_{(j)}) \right|^p \right]^{\frac{1}{p}}. \tag{39}$$

We now turn our attention to the optimization of the function ω parametrized in terms of $\boldsymbol{\alpha}$, $\omega(\cdot) = \sum_{i=1}^{m+n} \alpha_i \kappa(\cdot, z_i)$. For arbitrary sample sizes with $p = 2$, the max-sliced kernel Wasserstein-2 distance is

$$\textit{max-}W_2^{\mathcal{H}}(\hat{\mu}, \hat{\nu}) = \max_{\omega \in \mathcal{H}: \|\omega\|_2 \leq 1} \left(\|\omega\|_{\hat{\mu}}^2 + \|\omega\|_{\hat{\nu}}^2 - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \langle \mathbf{P}, \boldsymbol{\omega}_X \boldsymbol{\omega}_Y^\top \rangle \right)^{\frac{1}{2}}, \tag{40}$$

with unsorted values $\boldsymbol{\omega}_X = [\omega(x_1), \ldots, \omega(x_m)]^\top = \mathbf{K}_{XZ}\boldsymbol{\alpha}$ and $\boldsymbol{\omega}_Y = [\omega(y_1), \ldots, \omega(y_n)]^\top = \mathbf{K}_{YZ}\boldsymbol{\alpha}$.

The max-sliced kernel Wasserstein-2 distance can be expressed in terms of equation 37 or equation 40:

$$\begin{aligned}
\left[\textit{max-}W_2^{\mathcal{H}}(\hat{\mu}, \hat{\nu})\right]^2 &= \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \min_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \sum_{i=1,j=1}^{m,n} P_{ij} \left| \sum_{k=1}^{m+n} (\kappa(x_i, z_k) - \kappa(y_j, z_k))\alpha_k \right|^2 \\
&= \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \Big\{ \boldsymbol{\alpha}^\top \mathbf{K}_{XZ}^\top \mathbf{D}_\mu \mathbf{K}_{XZ} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{K}_{YZ}^\top \mathbf{D}_\nu \mathbf{K}_{YZ} \boldsymbol{\alpha} \\
&\qquad - 2 \max_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \boldsymbol{\alpha}^\top \mathbf{K}_{XZ}^\top \mathbf{P} \mathbf{K}_{YZ} \boldsymbol{\alpha} \Big\} \\
&= \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \min_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \boldsymbol{\alpha}^\top \mathbf{Q}_{\mathbf{P}} \boldsymbol{\alpha} = \max_{\boldsymbol{\alpha}} \frac{\min_{\mathbf{P} \in \mathcal{P}_{\hat{\mu},\hat{\nu}}} \boldsymbol{\alpha}^\top \mathbf{Q}_{\mathbf{P}} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}},
\end{aligned} \tag{41}$$

where $\mathbf{Q}_{\mathbf{P}} = \mathbf{K}_{XZ}^\top \mathbf{D}_\mu \mathbf{K}_{XZ} + \mathbf{K}_{YZ}^\top \mathbf{D}_\nu \mathbf{K}_{YZ} - \mathbf{K}_{XZ}^\top \mathbf{P} \mathbf{K}_{YZ} - \mathbf{K}_{YZ}^\top \mathbf{P}^\top \mathbf{K}_{XZ}$ is a symmetric matrix positive semidefinite matrix. (This can be seen since the objective is greater than or equal to zero for all choices of $\boldsymbol{\alpha}$.)

For fixed $\mathbf{P}$, equation 41 is a convex maximization, that can be solved as a generalized eigenvalue problem. For fixed $\boldsymbol{\alpha}$, the optimization in terms of $\mathbf{P}$ is a linear program with the solution detailed above. However, maximizing with respect to $\boldsymbol{\alpha}$, $\mathbf{P}$ is a matrix-valued function of $\boldsymbol{\alpha}$. To find a local maximum for $\boldsymbol{\alpha}$ we use the unconstrained optimization equation 41 and perform gradient ascent with respect to $\boldsymbol{\alpha}$, wherein each iteration we compute the optimal transport plan. This approach is also used in computing the generalized max-sliced Wasserstein distance Kolouri et al. (2019).

A.5 Covariate Shift Correction Algorithms

For covariate shift correction the goal is minimize the divergence by adjusting the weights $\boldsymbol{\nu}$ of the instances in one sample $\hat{\nu}$. This results in the following convex optimization:

$$\min_{\boldsymbol{\nu} \in \mathbb{R}_{\geq 0}^n : \langle \boldsymbol{\nu}, \mathbf{1} \rangle = 1} J(\boldsymbol{\nu}) \tag{42}$$

where

$$\begin{aligned}
J(\boldsymbol{\nu}) &= \left(\textit{max-}D_B^{\mathcal{H}}(\hat{\mu}, \hat{\nu})\right)^2 = \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \tilde{J}(\boldsymbol{\nu}, \boldsymbol{\alpha}) \\
&= \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \underbrace{\langle \boldsymbol{\mu}, (\mathbf{K}_{XZ}\boldsymbol{\alpha})^{\circ 2} \rangle}_{C_{\boldsymbol{\alpha}}} + \langle \boldsymbol{\nu}, \underbrace{(\mathbf{K}_{YZ}\boldsymbol{\alpha})^{\circ 2}}_{\mathbf{k}_{\boldsymbol{\alpha}}} \rangle - 2\sqrt{\langle \boldsymbol{\mu}, (\mathbf{K}_{XZ}\boldsymbol{\alpha})^{\circ 2} \rangle \langle \boldsymbol{\nu}, (\mathbf{K}_{YZ}\boldsymbol{\alpha})^{\circ 2} \rangle} \\
&= \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} C_{\boldsymbol{\alpha}} + \langle \boldsymbol{\nu}, \mathbf{k}_{\boldsymbol{\alpha}} \rangle - 2\sqrt{C_{\boldsymbol{\alpha}}} \sqrt{\langle \boldsymbol{\nu}, \mathbf{k}_{\boldsymbol{\alpha}} \rangle}.
\end{aligned}$$

For fixed $\boldsymbol{\alpha}$ the function $\tilde{J}(\cdot, \boldsymbol{\alpha})$ is a sum of a linear function and a convex function $f(\cdot) = -\sqrt{\cdot}$ is convex since $\sqrt{\cdot}$ is concave. $J(\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}: \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \leq 1} \tilde{J}(\boldsymbol{\nu}, \boldsymbol{\alpha})$ is convex since maximizing over $\boldsymbol{\alpha}$ preserves convexity (Nesterov, 2018, Theorem 3.1.8).

In the linear kernel case, the cost is

$$\begin{aligned}
J(\boldsymbol{\nu}) &= \left(\textit{max-}D_B^{\mathbb{R}^d}(\hat{\mu}, \hat{\nu})\right)^2 = \max_{\mathbf{w}: \|\mathbf{w}\|_2 \leq 1} \left\{ \tilde{J}(\nu, \mathbf{w}) = \left(\sqrt{\langle \boldsymbol{\mu}, (\mathbf{X}^\top \mathbf{w})^{\circ 2} \rangle} - \sqrt{\langle \boldsymbol{\nu}, (\mathbf{Y}^\top \mathbf{w})^{\circ 2} \rangle} \right)^2 \right\} \\
&= \max_{\mathbf{w}: \|\mathbf{w}\|_2 \leq 1} (\langle \boldsymbol{\mu}, (\mathbf{X}^\top \mathbf{w})^{\circ 2} \rangle + \langle \boldsymbol{\nu}, (\mathbf{Y}^\top \mathbf{w})^{\circ 2} \rangle - 2\sqrt{\langle \boldsymbol{\mu}, (\mathbf{X}^\top \mathbf{w})^{\circ 2} \rangle \langle \boldsymbol{\nu}, (\mathbf{Y}^\top \mathbf{w})^{\circ 2} \rangle}.
\end{aligned}$$

Again, $J(\boldsymbol{\nu})$ is convex since $\tilde{J}(\cdot, \mathbf{w})$ is convex for fixed $\mathbf{w}$. In this case, the gradient has the intuitive form

$$\nabla_{\boldsymbol{\nu}} J(\boldsymbol{\nu}) = \left(1 - \frac{\|\omega\|_{\hat{\mu}}}{\|\omega\|_{\hat{\nu}}}\right)(\boldsymbol{\omega}_{\mathbf{Y}}^{\top})^{\circ 2} = \left(1 - \frac{\sqrt{\langle \boldsymbol{\mu}, (\mathbf{X}^{\top}\mathbf{w}_{\boldsymbol{\nu}})^{\circ 2}\rangle}}{\sqrt{\langle \boldsymbol{\nu}, (\mathbf{Y}^{\top}\mathbf{w}_{\boldsymbol{\nu}})^{\circ 2}\rangle}}\right)(\mathbf{Y}^{\top}\mathbf{w}_{\boldsymbol{\nu}})^{\circ 2}. \tag{43}$$

To solve this convex minimization over a probability simplex we apply the Frank-Wolfe (conditional gradient) algorithm Jaggi (2013) to iteratively adjust the weight of one instance $\boldsymbol{\nu} \leftarrow (1-\gamma)\boldsymbol{\nu} + \gamma \mathbf{e}_i$, where $\mathbf{e}_i$ is an indicator vector, $i = \arg\min_{1 \leq j \leq n} [\nabla_{\boldsymbol{\nu}} J(\boldsymbol{\nu})]_j$, and $\gamma \in [0, 1]$ is the stepsize. A benefit of the Frank-Wolfe scheme is that it requires only rank-1 updates of $\mathbf{Y}\mathbf{D}_{\boldsymbol{\nu}}\mathbf{Y}^{\top}$, which are needed for updating $\mathbf{w}$.

A.6 Additional Experiments Results

Figure 6 details witness function evaluations for six data sets generated from two-dimensional distributions. We compare MMD and the kernel-based max-sliced Bures distance (MSB). Notably, the one-sided MSB divergences correspond to localized regions, which are not distributed outliers. This is beneficial for the 'precision' of the witness function, but the 'recall' of MMD is better. This behavior may or may not be beneficial depending on the task.

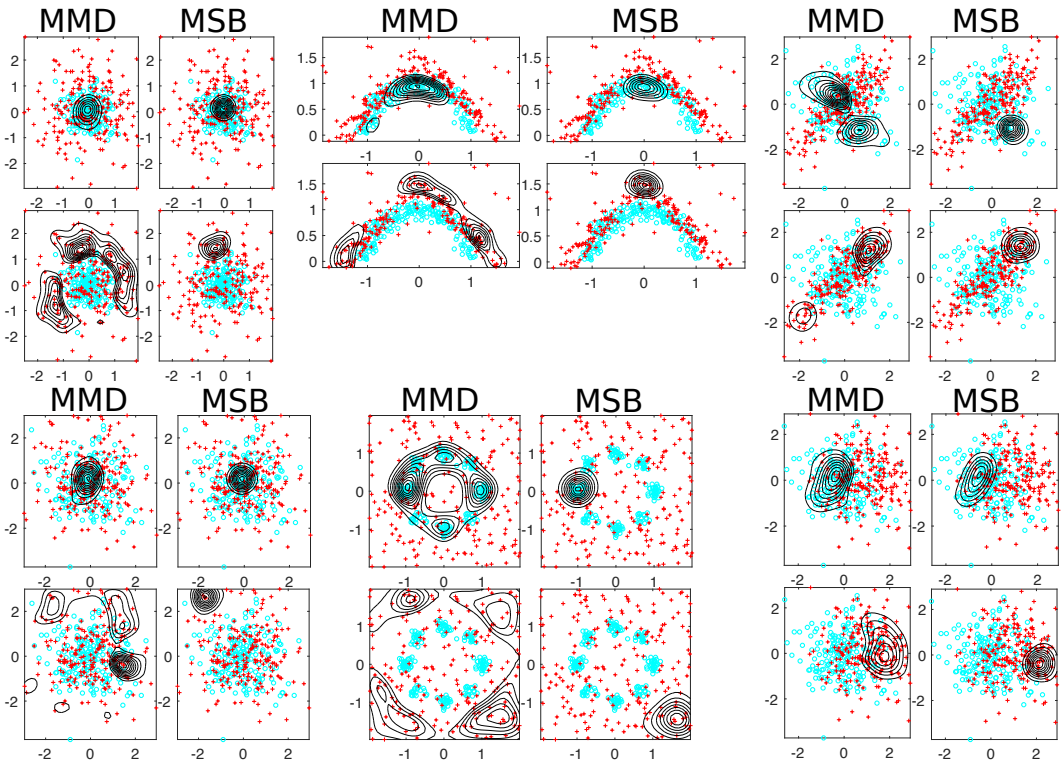

Figure 6: Maximum mean discrepancy (MMD) and max-sliced Bures distance (MSB) applied to two-dimensional samples using a Gaussian kernel. For each data set (shown as a two-by-two subplot), the contour plots indicate the squared magnitude of the witness function evaluations. For MMD, positive witness function values are plotted in the top row and negative evaluations are in the second row. For MSB, the rows correspond to the two one-sided divergences. The one-sided MSB divergences correspond to localized regions, which are not distributed outliers.

Figure 7 shows the results after 250 iterations of gradient descent with Adam (Kingma & Ba, 2015) to minimize the max-sliced Wasserstein-2 distance and max-sliced Bures distance by moving the points in one sample—originally from a uniform distribution—to match a mixture of 25 Gaussians. In two-dimensional space, the max-sliced Bures distance with a linear kernel fails. However, both divergences succeed (with a large number of outlier points left behind) after using a non-linear

mapping $\boldsymbol{\phi}(x)$ defined by $\phi(x)_i = \cos(w_{i1}x_1 + w_{i2}x_2 + c_i)$, where $w_{i1}, w_{i2} \sim \mathcal{N}(0, \sigma^2)$ and $c_i \sim U(0, 2\pi)$ $i \in \{1, \ldots, d\}$. After scaling $\langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(y) \rangle$ approximates a Gaussian kernel with kernel size σ (Rahimi & Recht, 2008).

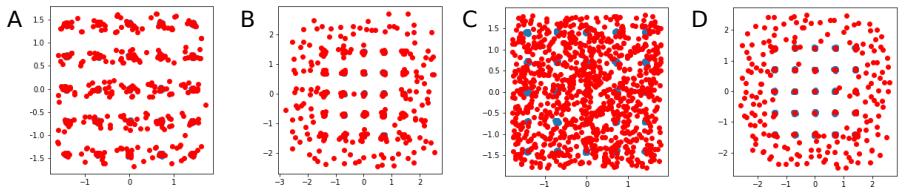

Figure 7: Optimizing the points originally drawn from a uniform distribution to match a mixture of 25 Gaussians by minimizing the (A,B) max-sliced Wasserstein-2 distance or (C,D) max-sliced Bures distance with a (A,C) linear kernel or (B,D) non-linear random Fourier bases with d=2,000 with σ=0.2. Example follows Kolouri et al. (2019) and uses Adam (Kingma & Ba, 2015) with default settings besides a learning rate of 10^{-2}. The max-sliced Wasserstein-2 distance succeeds with the original representation. Both distances are able to find the modes, but a large number of points are left behind.

Figure 8 shows 20 synthetic images—generated by AutoGAN trained on CIFAR10 (n=50,000)—with the highest weights after re-weighting in order to minimize the max-sliced Bures distance to the subset of training images for each class separately (m=5,000). The realism scores of the selected images have a median and range of 0.96 (0.63–1.34). Figure 9 shows the same but based on the weights optimized by using the Wasserstein-2 distance with the mini-batch optimization. The realism scores of the selected images have a median and range of 1.1 (0.93–1.29). Figure 10 shows the same but based on the weights optimized by using the max-sliced Wasserstein-2 distance with the mini-batch optimization. (In contrast, the max-sliced Bures distance is computationally tractable, since it does not depend on the sample size.) The realism scores of the selected images have a median and range of 0.97 (0.65–1.25).

The optimizations in all three cases use the Frank-Wolfe algorithm (Jaggi, 2013) with simplex constraints. The default step-size schedule $\gamma = \frac{2}{k+2}$ and the same stopping criterion is used $\max_{1 \leq i \leq n} |\nu_i^{(k)} - \nu_i^{(k-1)}| < 10^{-3}$. This yields roughly the same number of iterations for each method. The optimization starts from a uniform weighting, which means only $\sim$2000 instances are actually reweighted. The Fréchet inception distances after reweighting are detailed in Table 4. Based on the quantitative and qualitative results it appears that the Wasserstein-2 distance with mini-batch approximation weights high-quality synthetic images the highest, but the diversity of the highly weighted instances may not capture the full distribution for a class. In this regard, the max-sliced Bures better captures the diversity of the class, albeit choosing less realistic images.

Table 4: Fréchet inception distances (FID) between CIFAR10 test set images in each class and reweighted sample of synthetic images from AutoGAN. The second column shows the FID to the corresponding training set. The third column is a uniform weighting over all 50,000 synthetic images. The reweighting that minimizes the max-sliced Bures distance (MSB) to the subset of training images performs the best on average. Using the Wasserstein-2 distance (W2)—estimated through 10 mini-batches of 100 images on each iteration—performs best only on one-class. The max-sliced Wasserstein-2 (MSW2) distance also uses mini-batches.

	Training set	Uniform	MSB	W2	MSW2
airplane	28.00	108.67	**57.82**	74.67	81.85
automobile	20.20	133.10	**39.88**	66.76	74.81
bird	30.25	86.23	58.71	**57.17**	76.98
cat	35.53	84.63	**61.27**	83.89	75.57
deer	26.46	85.55	**46.81**	60.21	56.44
dog	28.99	106.76	**56.88**	67.94	78.77
frog	29.88	107.88	**51.14**	75.23	66.93
horse	24.33	111.30	**42.68**	66.77	63.24
ship	21.49	131.92	**37.35**	65.05	74.47
truck	17.77	141.56	**38.97**	63.73	76.05
Average	26.29	109.76	**49.15**	68.14	72.51

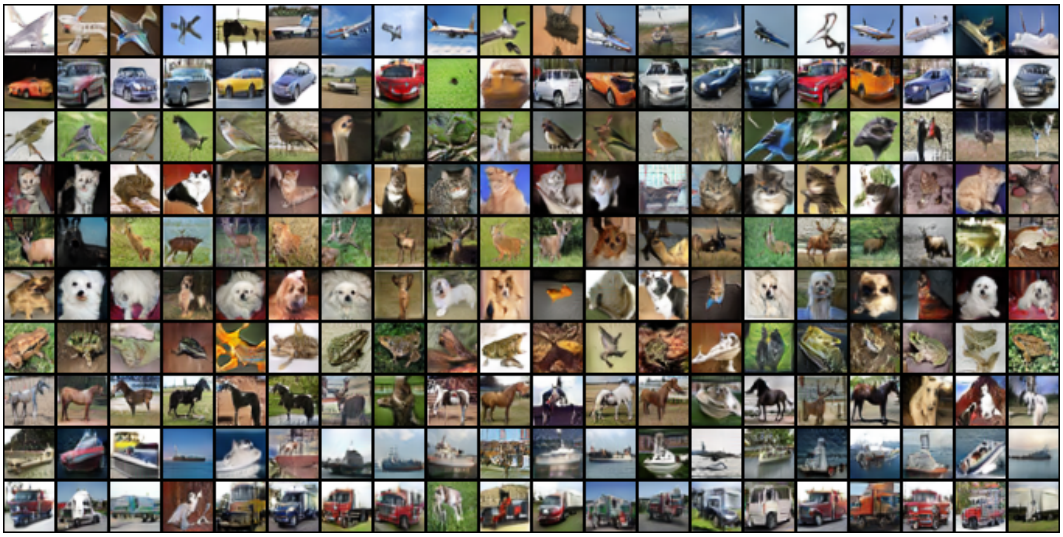

Figure 8: Distribution matching based on minimizing max-sliced Bures distance $\textit{max-}D_B^{\mathbb{R}^d}(\hat{\mu}, \hat{\nu})$. Synthetic images shown are those with the largest values of $\boldsymbol{\nu}$, where $\hat{\nu}$ is the $\boldsymbol{\nu}$-weighted distribution over 50,000 synthetic images from AutoGAN and $\hat{\mu}$ consists of CIFAR10 training images for a single class in each row. Rows (top to bottom) correspond to airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck classes.

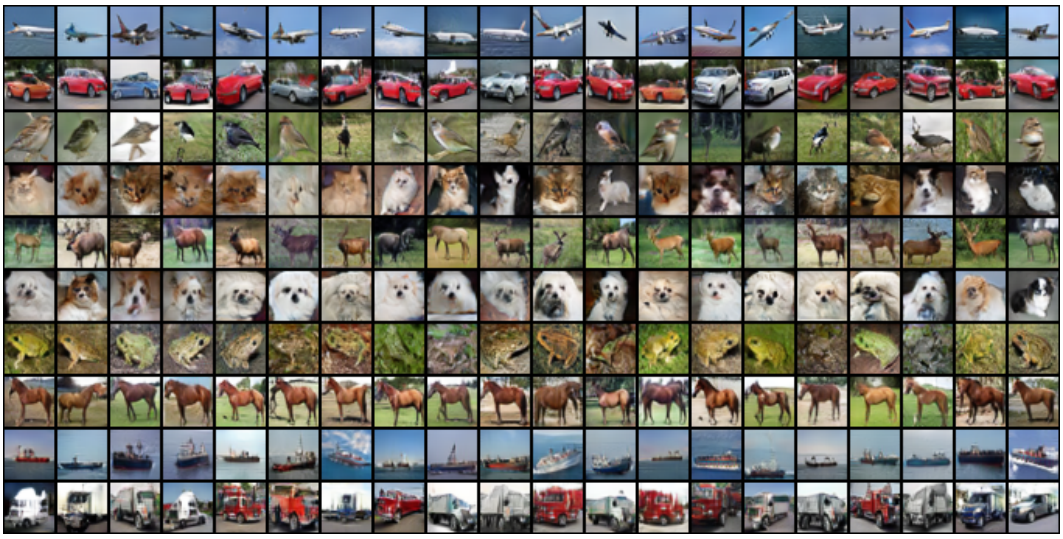

Figure 9: Distribution matching based on minimizing the Wasserstein-2 distance through mini-batch. Synthetic images shown are those with the largest values of $\boldsymbol{\nu}$, where $\hat{\nu}$ is the $\boldsymbol{\nu}$-weighted distribution over 50,000 synthetic images from AutoGAN and $\hat{\mu}$ consists of CIFAR10 training images for a single class in each row. Rows (top to bottom) correspond to airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck classes.

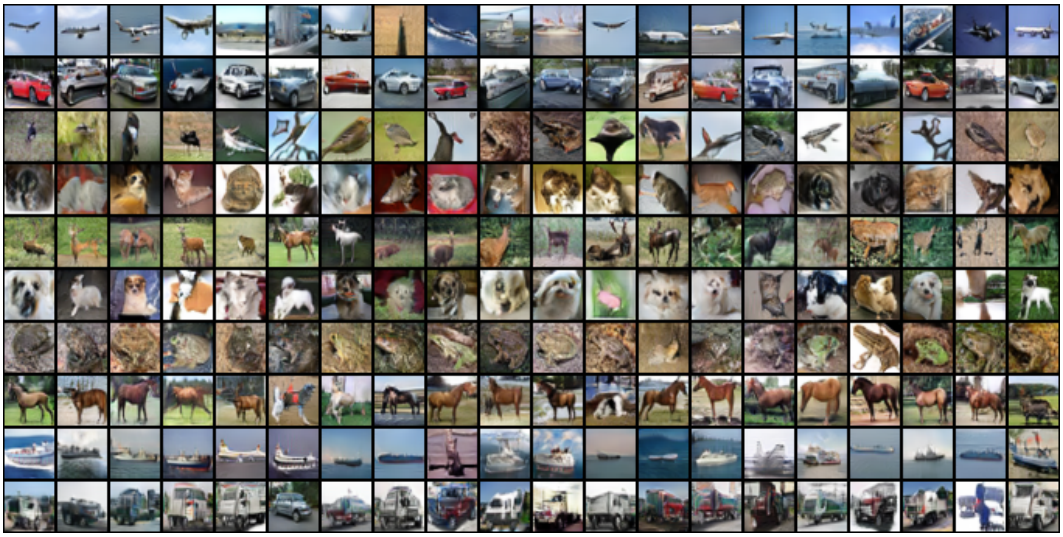

Figure 10: Distribution matching based on minimizing the max-sliced Wasserstein-2 distance *max*-$W_2^{\mathbb{R}^d}(\hat{\mu}, \hat{\nu})$ through mini-batch approximation. Synthetic images shown are those with the largest values of $\boldsymbol{\nu}$, where $\hat{\nu}$ is the $\boldsymbol{\nu}$-weighted distribution over 50,000 synthetic images from AutoGAN and $\hat{\mu}$ consists of CIFAR10 training images for a single class in each row. Rows (top to bottom) correspond to airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck classes.

Figure 11: Selecting images directly with the largest realism scores. Synthetic images shown are those with the largest realism values over 50,000 synthetic images generated by AutoGAN when the realism scores used in each row are computed using the CIFAR10 training images for a single class. Rows (top to bottom) correspond to airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck classes. Realism score correctly identifies "realistic" imagery, but is unable to find samples that cover the real distribution. For example, the top row is missing airplanes, the third row is missing birds, the fourth row is missing cats, etc.