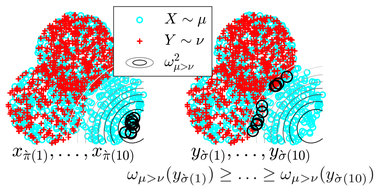
Identifying the Instances Associated with Distribution Shifts using the Max-Sliced Bures Divergence

Austin J. Brockmeier, University of Delaware
Claudio César Claros-Olivares, University of Delaware
Matthew S. Emigh, Naval Surface Warfare Center - Panama City Division
Luis Gonzalo Sanchez Giraldo, University of Kentucky

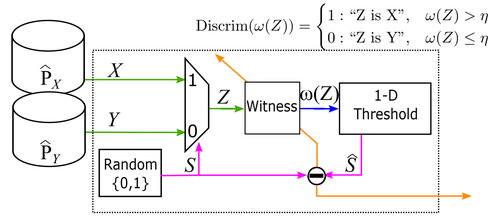
What is the goal?

To find examples of discrepancies between two data sets using **Interpretable statistical divergences**

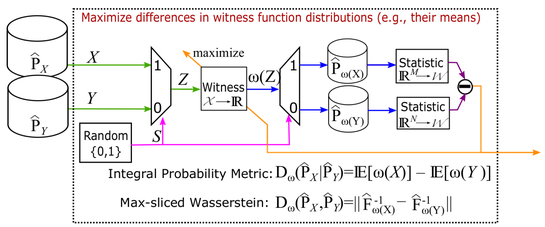
Interpret: examining the top-K examples from each sample from the witness function



Divergence as a Learning Problem



Maximal Discrepancy Divergences as a Learning Problem



Existing Maximal Discrepancy Divergences

- Maximum mean discrepancy (MMD)
  $\text{MMD}^{\mathcal{H}}(\mu,\nu) = \sup_{\omega \in \mathcal{F}} \mathbb{E}_{X\sim\mu, Y\sim\nu}[\langle \phi(X) - \phi(Y), \omega \rangle] = \sup_{\omega\in\mathcal{F}} \mathbb{E}[\omega(X) - \omega(Y)] = \|m_\mu - m_\nu\|_{\mathcal{H}}$
- Max-Sliced Wasserstein-2 (squared)
  - Saddlepoint optimization problem
  - Sample based O($N\log(N)$)

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

New Maximal Discrepancy Divergences

- Max-Sliced Bures (MSB) distance $\sup_{\omega\in\Omega} \sqrt{\mathbb{E}[\omega^2(X)]} - \sqrt{\mathbb{E}[\omega^2(Y)]}$
  - (One-sided) Max-Sliced Bures $\Omega = \{\omega(\cdot) = \langle \cdot, \mathbf{w}\rangle : \mathbf{w} \in \mathbb{R}^d, \quad \|\mathbf{w}\|_2 \leq 1\}$
  - (One-sided) Max-Sliced Kernel Bures $\Omega = \{\omega(\cdot) = \langle \phi(\cdot), \omega\rangle, \quad \omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} \leq 1\}$

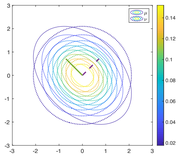
Wasserstein-2 distance between Gaussians is the Fréchet distance:
√(Squared Difference of Means + Squared Bures between Cov.)

Prove: Sliced Bures ≤ Sliced Fréchet ≤ Sliced Wasserstein-2

- Only detects differences in 1st or 2nd moments
- For higher moments, rely on pre-trained learning representation (Inception codes):
  - Interpret the Fréchet Inception Distance
- Kernel approach
  - MMD is the difference of means in RKHS
  - Exploit characteristic kernels, estimate witness function in RKHS
  - Scale with random Fourier features (Rahimi and Recht, 2007)

Max-Sliced Bures Divergence

$$\mathbb{E}[\langle X, \mathbf{w}\rangle^2] = \mathbf{w}^\top \mathbb{E}[XX^\top]\mathbf{w} = \mathbf{w}^\top \boldsymbol{\rho}_X \mathbf{w}$$
$$D_{\text{MSB}}(\mu,\nu) = \sup_{\mathbf{w}\in\mathcal{S}} \left| \sqrt{\mathbf{w}^\top \boldsymbol{\rho}_X \mathbf{w}} - \sqrt{\mathbf{w}^\top \boldsymbol{\rho}_Y \mathbf{w}} \right|$$
$$= \max\left\{ \sqrt{\mathbb{E}[\omega_{\mu>\nu}(X)]} - \sqrt{\mathbb{E}[\omega_{\mu>\nu}(Y)]}, \sqrt{\mathbb{E}[\omega_{\mu<\nu}(Y)]} - \sqrt{\mathbb{E}[\omega_{\mu<\nu}(X)]} \right\}$$



Optimal witness functions: $\omega_{\mu>\nu}(\cdot) = \langle \cdot, \mathbf{w}_{\mu>\nu}\rangle^2 \qquad \omega_{\mu<\nu}(\cdot) = \langle \cdot, \mathbf{w}_{\mu<\nu}\rangle^2$

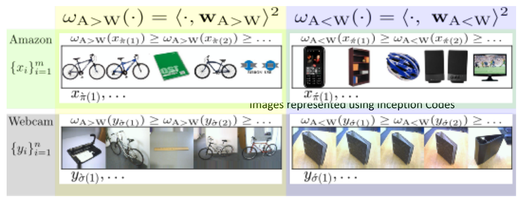
Alg.1: Find the max-slice for Bures is 1D bounded line search and primary eigenvector

Algorithm 1: One-sided max-sliced Bures divergence

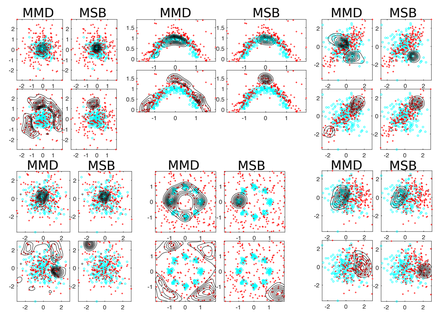
Input: $\boldsymbol{\rho}_X = \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i^\top, \boldsymbol{\rho}_Y = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top \in \mathbb{R}^{d\times d}, \epsilon > 0$
Define the function $\mathbf{v}_{(\cdot)} : \mathbb{R} \to \mathbb{R}^d$ as $\mathbf{v}_{(\gamma)} : \gamma \mapsto \arg\max_{\mathbf{w}:\|\mathbf{w}\|_2\leq 1} \mathbf{w}^\top(\gamma\boldsymbol{\rho}_X - \boldsymbol{\rho}_Y)\mathbf{w}$
Solve the 1D bound problem: $\gamma^\star = \arg\max_{0<\gamma\leq 1} \sqrt{\mathbf{v}_{(\gamma)}^\top \boldsymbol{\rho}_X \mathbf{v}_{(\gamma)}} - \sqrt{\mathbf{v}_{(\gamma)}^\top \boldsymbol{\rho}_Y \mathbf{v}_{(\gamma)}}$
Output: $\mathbf{w}_{\mu>\nu} = \mathbf{v}_{(\gamma^\star)}$

Alternative: Gradient approach using smoothed square roots

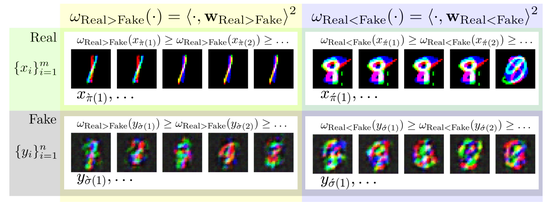
Detecting discrepancies in domain transfer



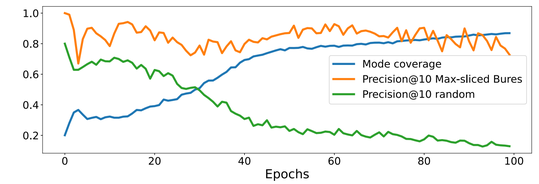
Kernelized Max-sliced Bures (MSB) is more localized than Maximal Mean Discrepancy (MMD)



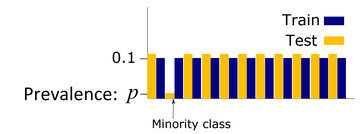
Interpreting GANs: examining the top-5 examples from each sample from each witness function



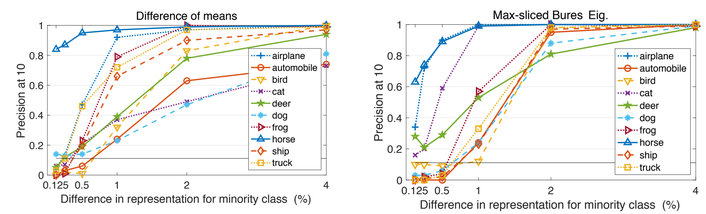
Precision of the witness function in detecting dropped modes



Precision of the witness function in detecting underrepresented classes

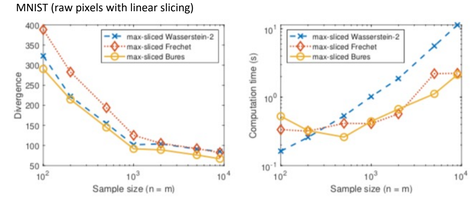


CIFAR10 using Inception Codes (linear slicing)



Max-sliced Fréchet via Max-sliced Bures is more accurate than saddlepoint optimization approach for max-sliced Wasserstein-2

MNIST (raw pixels with linear slicing)



Conclusions:

- Identify localized discrepancies through interpretable divergences
- Highlight opportunities for better dataset calibration
- Max-slicing the Bures distance is scalable and interpretable
- Dropped modes can be detected by looking at instances where the witness function has the largest magnitude
- Class imbalances can be recognized efficiently