Adversarial Support Alignment via Relaxed 1D Optimal Transport

Shangyuan Tong [*1] **Timur Garipov** [*1] **Yang Zhang** [2] **Shiyu Chang** [2] **Tommi Jaakkola** [1]

Abstract

Distribution alignment has become a broadly useful subroutine, from generative models to domain adaptation. However, strict distribution alignment is rarely achieved nor desired. For example, recent work has shown that a simple shift in label distribution between domains would already introduce undesirable distortions under strict distribution alignment of associated latent representations. Instead, in this paper we focus on and develop methods for aligning only the distribution support. Our work builds on easily solvable relaxed optimal transport problems between 1D pushforward measures under discriminator mapping. We show, for example, that the signal from the standard log-loss discriminator used in adversarial distribution alignment can indicate support discrepancy while other discriminator choices may fail. We establish a connection between a family of relaxed 1D optimal transportation problems and a spectrum of metrics corresponding to different notions of alignment. In particular, we compute the support divergence via solving a symmetrized ∞-to-1 optimal transportation assignment. We provide theoretical and empirical results that illustrate the properties and impact of the proposed support alignment framework.

1. Introduction

Developing methods for distribution alignment has become an active area of research. In generative modeling, for example, distribution alignment is essentially the end goal. Given a fixed real data distribution p_r, we aim to learn a distribution p_g such that $p_g \approx p_r$. Many popular methods are based on Generative Adversarial Networks (Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017; Radford et al., 2015). The distributions in GANs are specified implicitly, as push-forward measures from a simpler latent space to observations, and learned by distributional matching. Distribution alignment plays an equally prevalent role in domain adaptation. The goal in this case is expressed in terms of a feature extractor, and measured between the latent feature representations corresponding to examples from the source and target domains. Motivated by theoretical results (Ben-David et al., 2007; 2010), a series of papers (Ajakan et al., 2014; Ganin & Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2017; Shen et al., 2018; Pei et al., 2018; Zhao et al., 2018; Li et al., 2018a; Wang et al., 2021; Kumar et al., 2018) seek to align representations between domains in a variety of contexts.

Perfect distribution alignment is not easy to accomplish in high-dimensions, nor is it necessarily desirable. For generative models, for example, distribution alignment implies memorizing proportional "biases" in the dataset. For example, if the training images are composed of 80% dogs and 20% cats, then the resulting generative model reproduces this frequency bias. In domain adaptation, recent works (Zhao et al., 2019; Li et al., 2020; Tan et al., 2020; Wu et al., 2019b; Tachet des Combes et al., 2020) have demonstrated that a shift in label distributions between the source and target examples leads to a characterizable performance hit if the domains are forced into a distribution alignment. The induced distortions from alignment would naturally affect (unlabeled) subgroups similarly.

In response to the concerns with strict alignment, Wu et al. (2019b) proposed to relax the distribution alignment constraint. They allowed the density ratio (in its defined region) to be upper-bounded by some fixed constant rather than enforcing distributions to agree exactly everywhere. They provided both theoretical and empirical arguments in favor of relaxing the alignment.

In this paper, we take relaxation a step further, and develop methods to align only the distribution support. Support alignment is an easier constraint to satisfy than distribution alignment. Indeed, distribution alignment implies support alignment but not vice versa. Support alignment may also match the stated goal better. For generative models, intuitively, only the support needs to align with real examples in order for the model to produce realistic samples. In terms

[*]Equal contribution [1]MIT Computer Science & Artificial Intelligence Lab, Cambridge, MA, USA [2]MIT-IBM Watson AI Lab, Cambridge, MA, USA. Correspondence to: Shangyuan Tong <sytong@csail.mit.edu>, Timur Garipov <timur@csail.mit.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.

of domain adaptation, Johansson et al. (2019) analyze various failure cases of distribution alignment, and also suggest aligning the support to avoid unwanted distortions. We position support alignment within a spectrum of alignment objectives and relate them theoretically. A broader scope of the related work is discussed in Appendix A.

2. SSD divergence and spectrum of alignment

Notation We consider a Euclidian space $\mathcal{X} = \mathbb{R}^n$ equipped with Borel sigma algebra $\mathcal{B}$ and a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{P}$ be the set of probability measures on $(\mathcal{X}, \mathcal{B})$. For a probability measure $p \in \mathcal{P}$ the support of p is denoted by $\mathrm{supp}(p)$ and is defined as the smallest closed set $X_p \subseteq \mathcal{X}$ such that $p(X_p) = 1$. The distance between a point $x \in \mathcal{X}$ and a subset $Y \subseteq \mathcal{X}$ is defined as $d(x, Y) = \inf_{y \in Y} d(x, y)$.

2.1. SSD divergence

In order to achieve support alignment, we first need to evaluate how well the supports are aligned. Similar to distribution divergence like Kullback–Leibler divergence and Wasserstein distance, we provide an intuitive definition for support divergence. A *support divergence*[1] between two distributions in $\mathcal{P}$ is a function $\mathcal{D}_S(\cdot, \cdot) : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ satisfying: (1) $\mathcal{D}_S(p, q) \geq 0$ for all $p, q \in \mathcal{P}$; (2) $\mathcal{D}_S(p, q) = 0$ if and only if $\mathrm{supp}(p) = \mathrm{supp}(q)$.

While a distribution divergence is sensitive to both density and support discrepancies, a support divergence only needs to detect differences in supports, which are subsets of the metric space $\mathcal{X}$. An example of a distance between subsets of a metric space is the well-known Hausdorff distance. Adapting the Hausdorff distance to distributions, we introduce *symmetric support difference divergence* (*SSD divergence*)[2]:

$$\mathcal{D}_\triangle(p, q) = \mathbb{E}_p \left[d(x, \mathrm{supp}(q))\right] + \mathbb{E}_q \left[d(x, \mathrm{supp}(p))\right]. \quad (1)$$

Proposition 2.1. *SSD divergence is a support divergence.*

2.2. Spectrum of alignment

Wu et al. (2019b) proposed a modified Wasserstein distance to achieve asymmetrically-relaxed distribution alignment, namely β-admissible Wasserstein distance:

$$\mathcal{D}_W^\beta(p, q) = \inf_{\gamma \in \Gamma_\beta(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)], \quad (2)$$

where $\Gamma_\beta(p, q)$ is the set of all measures γ on $\mathcal{X} \times \mathcal{X}$ such that $\int \gamma(x, y)dy = p(x), \forall x$ and $\int \gamma(x, y)dx \leq (1 + \beta)q(y), \forall y$. They also showed that $\mathcal{D}_W^\beta(p, q) = 0$ if and only if $\sup_{x\in\mathcal{X}} p(x)/q(x) \leq 1 + \beta$.

We can extend the β-admissible Wasserstein distance to a symmetric version, which we term β_1, β_2-admissible Wasserstein distance:

$$\mathcal{D}_W^{\beta_1,\beta_2}(p, q) = \mathcal{D}_W^{\beta_1}(p, q) + \mathcal{D}_W^{\beta_2}(q, p). \quad (3)$$

The aforementioned property of β-admissible Wasserstein distance implies that $\mathcal{D}_W^{\beta_1,\beta_2}(p, q) = 0$ if and only if $(1 + \beta_2)^{-1} \leq p(x)/q(x) \leq 1 + \beta_1, \forall x \in \mathrm{supp}(p) \cup \mathrm{supp}(q)$, where we call p and q "(β_1, β_2)-aligned".

In the extreme case $\beta_1 = \beta_2 = 0$, $\mathcal{D}_W^{0,0}(p, q) = 2\mathcal{D}_W(p, q)$, where $\mathcal{D}_W(p, q)$ is the Wasserstein-1 distance with transportation cost $d(\cdot, \cdot)$. Now, we consider the limit in the opposite direction, i.e. $\mathcal{D}_W^{\infty,\infty}(p, q) \coloneqq \lim_{\beta\to\infty} \mathcal{D}_W^{\beta,\beta}(p, q)$.

Proposition 2.2. $\mathcal{D}_W^{\infty,\infty}(p, q) = \mathcal{D}_\triangle(p, q)$.

The following proposition establishes a relationship within the spectrum of alignment objectives.

Proposition 2.3. *For any p, q and for any finite $\beta_1, \beta_2 > 0$,*

1. $\mathcal{D}_W(p, q) = 0 \Rightarrow \mathcal{D}_W^{\beta_1,\beta_2}(p, q) = 0$.
2. $\mathcal{D}_W^{\beta_1,\beta_2}(p, q) = 0 \Rightarrow \mathcal{D}_\triangle(p, q) = 0$.
3. *The converse of statements 1 and 2 are false.*

3. Quantifying alignment in 1D spaces

Recent works (Rabin et al., 2011) have proposed to evaluate the distribution divergence by reducing the OT problem in the original, potentially high-dimensional space, to that between 1D distributions. Specifically, they use the sliced Wasserstein distance — the average of Wassertein distances between 1D pushforward distributions induced by linear projections on all possible directions. Sliced Wasserstein distance is proposed for distribution alignment and its use is justified by the fact that for any $p \neq q$ there exists a linear function $f^\theta(x) = \langle\theta, x\rangle$, which identifies difference in the distributions, i.e. ${f^\theta}_\sharp p \neq {f^\theta}_\sharp q$ (Cramér–Wold theorem).

Unfortunately, we find that support alignment of all linear projections of given distributions does not guarantee alignment of supports in the original space. In other words the family of linear mappings $f^\theta(x) = \langle\theta, x\rangle$ can not always identify the difference between supports of distributions (see Appendix B.2 for details). In the next subsection we show that the support misalignment can be identified by a learned non-linear 1D function.

3.1. Alignment via log-loss discriminator

Goodfellow et al. (2014) show that the log-loss discriminator $f : \mathcal{X} \to [0, 1]$, trained to distinguish samples from distributions p and q ($\sup_f \mathbb{E}_{x\sim p} [\log f(x)] +$

[1] Note that, strictly speaking, support divergence is not a divergence on the space of probability distributions, since $\mathcal{D}_S(p, q) = 0$ does not imply $p = q$.

[2] We adopt the set-theoretic operation notation: "$\triangle$" represents the symmetric set difference.

$\mathbb{E}_{y\sim q}[\log(1-f(y))]$) can be used to estimate the Jensen-Shannon divergence between p and q. The closed form maximizer f^* is

$$f^*(x) = \frac{p(x)}{p(x)+q(x)}, \qquad \forall x \in \text{supp}(p) \cup \text{supp}(q). \quad (4)$$

Using expression (4), we prove the following theorem characterizing the ability of discriminator to identify all three notions of misalignment discussed in Section 2.2.

Theorem 3.1. *Let f^* be the optimal discriminator (4) for given distributions p and q. Then,*

1. $\mathcal{D}_W(p,q) = 0 \Leftrightarrow \mathcal{D}_W(f^*{}_\sharp p, f^*{}_\sharp q) = 0$,
2. $\mathcal{D}_W^{\beta_1,\beta_2}(p,q) = 0 \Leftrightarrow \mathcal{D}_W^{\beta_1,\beta_2}(f^*{}_\sharp p, f^*{}_\sharp q) = 0$,
3. $\mathcal{D}_\triangle(p,q) = 0 \Leftrightarrow \mathcal{D}_\triangle(f^*{}_\sharp p, f^*{}_\sharp q) = 0$.

Remark 3.1.1. We note that in practice the discriminator is typically parameterized as $f(x) = \sigma(g(x))$, where $g : \mathcal{X} \to \mathbb{R}$ is realized by a deep neural network and $\sigma(x) = (1+e^{-x})^{-1}$ is the sigmoid function. The optimization problem for g is

$$\inf_g \ \mathbb{E}_p\left[\log(1+e^{-g(x)})\right] + \mathbb{E}_q\left[\log(1+e^{g(y)})\right], \quad (5)$$

and the optimal solution is $g^*(x) = \log p(x) - \log q(x)$. The result of Theorem 3.1 also holds for discriminator g^*.

3.2. 1D optimal transport with assignment constraints

Inspired by the result of Theorem 3.1, we develop methods to achieve different notions of alignment by minimizing different objectives (Section 2.2) between the 1D pushforward distributions. Because in practice, alignment-based methods typically operate with discrete samples, we consider 1D empirical distributions $\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(x)$, $\widehat{q}_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}(x)$ on a real line $x_i, y_i \in \mathbb{R}$. We use $\overline{1,n}$ to denote the set $\{1,\dots,n\} \subset \mathbb{N}$.

In section 2.2 we discussed the family of relaxed OT distances. Now, we adapt[3] this family of OT problems for empirical distributions $\widehat{p}_n$ and $\widehat{q}_n$:

$$\widehat{\mathcal{D}}_W^c(\widehat{p}_n, \widehat{q}_n) = \min_{\pi \in \Pi_c(\widehat{p}_n, \widehat{q}_n)} \frac{1}{n}\sum_{i=1}^n d(x_i, y_{\pi(i)}), \quad (6)$$

where $c \in \mathbb{N}$, $\Pi_c(\widehat{p}_n, \widehat{q}_n)$ is the set of mappings $\pi : \overline{1,n} \to \overline{1,n}$ such that $\#\{i \mid \pi(i) = j\} \le c, \forall j \in \overline{1,n}$.

The optimization goal in (6) is to find a "hard-assignment" π which describes the transportation of points x_i in $\widehat{p}_n$ to points y_j in $\widehat{q}_n$. The parameter c controls the set of admissible assignments Π_c. The constraints imposed on the assignments π are similar to the constraints on β-admissible coupling $\gamma \in \Gamma_\beta$ in (2): under a c-admissible assignment π, the total number of points x_i transported to each of the points y_j can not exceed c.

Below we consider three variants of the problem (6) related to the three notions of alignment discussed in Section 2.2.

Distribution alignment via 1-to-1 assignment Observe that $\widehat{\mathcal{D}}_W^1(\widehat{p}_n, \widehat{q}_n) = 0$ if and only if there exists an invertible (1-to-1) mapping π^*: $x_i = y_{\pi(i)}$ which ensures $\widehat{p}_n = \widehat{q}_n$.

Relaxed distribution alignment via c-to-1 assignment When $c > 1$, the assignment π in (6) can assign at most c points x_i to a single point y_j. We refer to such assignment as c-to-1. In order to draw the connection between the c-to-1 assignment and relaxed distribution alignment, we consider a symmetrized version of (6): $\widehat{\mathcal{D}}_W^{c_1,c_2}(\widehat{p}_n, \widehat{q}_n) = \widehat{\mathcal{D}}_W^{c_1}(\widehat{p}_n, \widehat{q}_n) + \widehat{\mathcal{D}}_W^{c_2}(\widehat{q}_n, \widehat{p}_n)$ defined by analogy with a symmetrized OT objective $\mathcal{D}_W^{\beta_1,\beta_2}$ in (3). Comparing (2) and (6), it is easy to see that $\widehat{\mathcal{D}}_W^{c_1,c_2}(\widehat{p}_n, \widehat{q}_n) = 0$ if and only if $\widehat{p}_n$ and $\widehat{q}_n$ are (β_1,β_2)-aligned with $\beta_1 = c_1 - 1$ and $\beta_2 = c_2 - 1$.

Support alignment via ∞-to-1 assignment When $c = \infty$, there are no constraints on assignment π in (6). Each point x_i can be assigned to any of the points y_j, or in terms of distributions, probability mass in $\widehat{p}_n$ can be transported to any location in the support of $\widehat{q}_n$. $\widehat{\mathcal{D}}_W^{\infty,\infty}(\widehat{p}_n, \widehat{q}_n) = 0$ if and only if $\text{supp}(\widehat{p}_n) = \text{supp}(\widehat{q}_n)$. Without assignment constraints in (6), the optimal transportation of $\widehat{p}_n$ to $\text{supp}(\widehat{q}_n)$ is realized by simply moving each x_i to its closest neighbor in $y_1, \dots y_n$.

4. Adversarial support alignment

In Section 3, we have seen that in the case of 1D empirical distributions the optimal transport problems discussed in Section 2.2 are reduced to easily solvable point-to-point assignment problems. Combining this observation with the result of Theorem 3.1, we propose a practical method for support alignment which we call *Adversarial Support Alignment* (*ASA*).

We work with distributions p_θ and q_θ parameterized by a vector of parameters θ. Our goal is to align the supports of p_θ and q_θ, i.e. to find θ^* such that $\text{supp}(p_{\theta^*}) = \text{supp}(q_{\theta^*})$. Motivated by Theorem 3.1 (and Remark 3.1.1) we train a log-loss discriminator g_ϕ, which is parameterized by ϕ. Given two mini-batches of samples $(x_1,\dots,x_n) \sim p_\theta$ and $(y_1,\dots,y_n) \sim q_\theta$ the discriminator gradient update is simply the gradient descent update on the log-loss (5).

While in adversarial distribution alignment methods (Good-

[3] Please see Appendix D for discussion of the connection between the "soft-assignment" OT problems discussed in 2.2 and the "hard-assignment" problem (6).

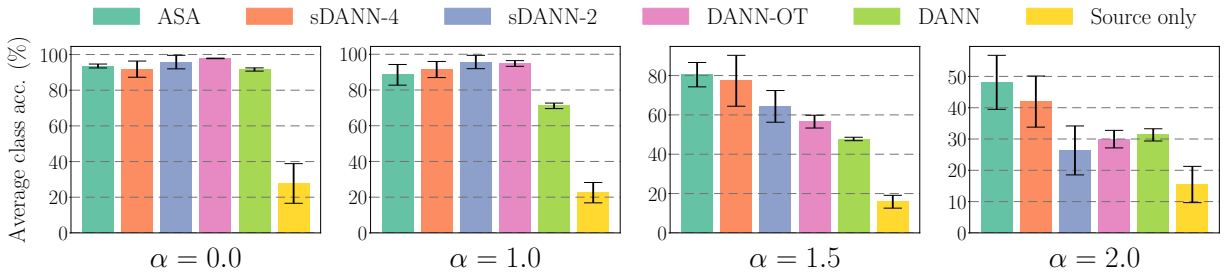

Figure 1. Comparison of alignment methods for domain adaption on MNIST-to-USPS task with class imbalance. The 4 panels shows the average class accuracy in the target domain for different severity level of class imbalance between source and target. For each method we show the mean accuracy and ± 1 standard deviation errorbars over 5 runs.

fellow et al., 2014; Ganin et al., 2016) θ competes with ϕ in a zero-sum game and simply maximizes the discriminator's objective, in ASA, θ minimizes the symmetrized optimal transportation cost $\widehat{\mathcal{D}}_W^{\infty,\infty}$ in the discriminator output space as described in Section 3.2. Recall that $\widehat{\mathcal{D}}_W^{\infty,\infty}$ is defined via two optimal ∞-to-1 assignments $\pi_{p\to q}$ and $\pi_{q\to p}$ matching point sets $\{g_\phi(x_i)\}_{i=1}^n$ to $\{g_\phi(y_i)\}_{i=1}^n$ in both directions. The mini-batch gradient w.r.t. θ is given by

$$\nabla_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \Big[d\left(g_\phi(x_i), g_\phi(y_{\pi_{p\to q}(i)})\right) + d\left(g_\phi(y_i), g_\phi(x_{\pi_{q\to p}(i)})\right) \Big] \right). \quad (7)$$

Effect of mini-batch training When training models on large datasets one has to rely on stochastic optimization with mini-batches. By using the gradient (7) the described algorithm minimizes the mini-batch transportaion cost rather than the population transportation cost. We observe that in practice this procedure enforces support alignment for all possible pairs of mini-batches which brings the distributions to the state closer to distribution alignment rather than support alignment (see Section E.1).

To address the mini-batch training issue, without having to solve the 1D assignment with full datasets, we introduce "history buffers" which store 1D discriminator output values from the several last mini-batches. For a new pair of mini-batches $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, instead of the mini-batch assignments as in (7), we find "mini-batch + history" assignments $\pi_{p\to q}^h$ and $\pi_{q\to p}^h$ between the combined "mini-batch + history" point sets. We provide a detailed description of this procedure in Appendix E.

5. Experiments

We present experimental evaluation of the proposed ASA method in domain adaptation setting.

We compare the following domain adaptation methods: (a) Source only training, (b) DANN (Ganin et al., 2016), (c) sDANN-β (Wu et al., 2019b), (d) DANN-OT (distribution alignment using 1D OT transport distance on the discriminator output) (e) ASA (our proposed method for support alignment). For DANN, sDANN and ASA, the classification model is composed of two parts, a feature extractor and a linear classifier which operates on the latent representation of the feature extractor. As discussed in Section 4 the considered methods aim to achieve distribution/support alignment between latent representation in source and target using the signal from a discriminator.

We consider the MNIST (LeCun et al., 1998) to USPS (Hull, 1994) adaptation task. We introduce class imbalance via reweighted sampling of training examples with the manually specified class distributions for source and target. We define the class priors $p_S(y)$ and $p_T(y)$ for source and target as power law distributions $p_S(y) \propto y^{-\alpha}$, $p_T(y) \propto (C - y + 1)^{-\alpha}$ where $y \in \{1, \dots, C\}$, C is the number of classes, and $\alpha \geq 0$ is the parameter controlling the imbalance. For $\alpha = 0$ both p_S and p_T give the uniform class prior, for $\alpha > 0$ the class proportions on source and target are imbalanced and the majority class in source corresponds to the minority class in target. In this class-imbalanced setting, we choose the target class average accuracy as the quality criterion for model evaluation. We use a 4-layer CNN as a feature extractor. For ASA we use history buffers of size 1000.

The results of all methods in the class-imbalanced setting are shown on Figure 1 and in Table F.1. Without any alignment, Source only training struggles to adapt to the target domain in this setting. Distribution alignment based methods (DANN, DANN-OT) perform well in the balanced setting, but suffer greatly otherwise. On the other hand, relaxed distribution alignment (sDANN-β) shows some resilience to this effect, but they have greater noise across their runs. Finally, support alignment (ASA) is the most robust method out of all, while still being competitive under desired settings for other methods (balanced for distribution alignment, and slight imbalance for relaxed distribution alignment).

Acknowledgements

The computational experiments presented in this paper were performed on "Satori" cluster developed as a collaboration between MIT and IBM. TJ acknowledges support from MIT-IBM Watson AI Lab and from Singapore DSO.

We thank Pavel Izmailov for helpful comments.

References

Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., and Marchand, M. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Balaji, Y., Chellappa, R., and Feizi, S. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems Foundation (NeurIPS)*, 2020.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bertsimas, D. and Tsitsiklis, J. N. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Budish, E., Che, Y.-K., Kojima, F., and Milgrom, P. Implementing random assignments: A generalization of the birkhoff-von neumann theorem. In *2009 Cowles Summer Conference*, 2009.

Chalapathy, R., Menon, A. K., and Chawla, S. Robust, deep and inductive anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36–51. Springer, 2017.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., and Kloft, M. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 3–17. Springer, 2018.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5769–5779, 2017.

Hoffmann, H. Kernel pca for novelty detection. *Pattern recognition*, 40(3):863–874, 2007.

Hull, J. J. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Johansson, F. D., Sontag, D., and Ranganath, R. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

Knorr, E. M., Ng, R. T., and Tucakov, V. Distance-based outliers: algorithms and applications. *The VLDB Journal*, 8(3):237–253, 2000.

Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., and Wornell, G. Co-regularized alignment for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 31:9345–9356, 2018.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, B., Wang, Y., Che, T., Zhang, S., Zhao, S., Xu, P., Zhou, W., Bengio, Y., and Keutzer, K. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018a.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018b.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Pei, Z., Cao, Z., Long, M., and Wang, J. Multi-adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Perera, P., Nallapati, R., and Xiang, B. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkQkBnJAb.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Sun, B., Feng, J., and Saenko, K. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Tachet des Combes, R., Zhao, H., Wang, Y.-X., and Gordon, G. J. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33, 2020.

Tan, S., Peng, X., and Saenko, K. Class-imbalanced domain adaptation: An empirical odyssey. In *European Conference on Computer Vision*, pp. 585–602. Springer, 2020.

Tax, D. M. and Duin, R. P. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Wang, J., Chen, J., Lin, J., Sigal, L., and de Silva, C. W. Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by gaussian-guided latent alignment. *Pattern Recognition*, 116:107943, 2021.

Wu, J., Huang, Z., Acharya, D., Li, W., Thoma, J., Paudel, D. P., and Gool, L. V. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019a.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pp. 6872–6881. PMLR, 2019b.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, pp. 727–736. IEEE, 2018.

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 8568–8579, 2018.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

A. Related work

This paper has connections with existing work in three directions: support estimation, distribution alignment, and relaxed distribution alignment.

Support estimation. There exists a series of work (e.g. (Schölkopf et al., 2001; Hoffmann, 2007; Tax & Duin, 2004; Knorr et al., 2000; Chalapathy et al., 2017; Ruff et al., 2018; Perera et al., 2019; Deecke et al., 2018; Zenati et al., 2018)) on novelty/anomaly detection problem, which can be casted as support estimation. We consider a fundamentally different problem setting. Our goal is to align the supports and our approach does not directly estimate the supports, instead, we implicitly learn the relationships between supports (density ratio to be specific) through a discriminator.

Distribution alignment. Apart from the works (e.g. (Ajakan et al., 2014; Ganin et al., 2016; Ganin & Lempitsky, 2015; Pei et al., 2018; Zhao et al., 2018; Long et al., 2018; Tachet des Combes et al., 2020; Li et al., 2018b; Tzeng et al., 2017; Shen et al., 2018; Kumar et al., 2018; Li et al., 2018a; Wang et al., 2021; Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017; Mao et al., 2017; Radford et al., 2015; Salimans et al., 2018; Genevay et al., 2018; Wu et al., 2019a; Deshpande et al., 2018; 2019)) that do distribution alignment, there are also papers (Long et al., 2015; 2017; Peng et al., 2019; Sun et al., 2016; Sun & Saenko, 2016) focusing on aligning some characteristics of the distribution, like first or second moment. Our work is concerned with a different problem, support alignment, which is a novel objective in this line of work. In terms of methodology, our use of the discriminator output space to work with easier optimizations in 1D is inspired by a line of work (Salimans et al., 2018; Genevay et al., 2018; Wu et al., 2019a; Deshpande et al., 2018; 2019) on sliced Wasserstein distance based models. Our discussion on such space also endorses the practical effectiveness of doing 1D OT for GAN in this space (Deshpande et al., 2019).

Relaxed distribution alignment. In the previous sections, we have already covered in detail about the connections between our work and that of Wu et al. (2019b). Balaji et al. (2020) also discuss a version of relaxed distribution alignment, but their focus is to align distributions but robust to outliers. There can be an interesting combination of this and our work, because our current algorithm considers all samples, which makes it not robust to outliers. We leave the possibility of a robust support alignment to future work.

B. Complimentary theoretical results

B.1. SSD divergence

Note that each term in (1) describes an asymmetric support difference:

$$\mathrm{SD}(p, q) = \mathbb{E}_{x\sim p}\left[d(x, \mathrm{supp}(q))\right] = \int_{\mathrm{supp}(p)\setminus\mathrm{supp}(q)} d(x, \mathrm{supp}(q))p(dx). \tag{8}$$

Thus, we can write

$$\mathcal{D}_{\triangle}(p, q) = \mathrm{SD}(p, q) + \mathrm{SD}(q, p).$$

B.2. Linear projections for support alignment

Recall that sliced Wasserstein distance (Rabin et al., 2011) is:

$$\widetilde{\mathcal{D}}_W(p, q) = \int_{\mathbb{S}^{n-1}} \mathcal{D}_W({f^\theta}_\sharp p, {f^\theta}_\sharp q)\, d\theta,$$

where f^θ is a 1D linear projection $f^\theta(x) = \langle\theta, x\rangle$, and $\mathbb{S}^{n-1} = \{\theta \in \mathbb{R}^n \mid \|\theta\| = 1\}$ is a unit sphere in $\mathbb{R}^n$.

As we mentioned in Section 3, the family of linear mappings $f^\theta(x) = \langle\theta, x\rangle$ can not always identify the difference between supports of distributions. The following proposition formalizes this claim.

Proposition B.1. *There exist two distributions p and q in $\mathcal{P}$, such that* $\mathrm{supp}(p) \neq \mathrm{supp}(q)$ *but* $\mathrm{supp}({f^\theta}_\sharp p) = \mathrm{supp}({f^\theta}_\sharp q)$, $\forall f^\theta(x) = \langle\theta, x\rangle$ *with* $\theta \in \mathbb{S}^{n-1}$.

B.3. Alignment via log-loss discriminator

First, note that for f^* as defined in (4), f^* is correctly defined in $\mathrm{supp}(p) \triangle \mathrm{supp}(q)$, e.g. for $x : p(x) > 0, q(x) = 0$ we have $p(x) + q(x) > 0$ and $f^*(x) = 1$. For a point $x \notin \mathrm{supp}(p) \cup \mathrm{supp}(q)$ the value of $f^*(x)$ can be set to an arbitrary value in $[0, 1]$, since the log-loss does not depend on $f(x)$ for such x.

Using (4), we can establish a connection between the pushforward distributions ${f^*}_\sharp p$ and ${f^*}_\sharp q$.

Proposition B.2. *Let f^* be the optimal discriminator (4) for given distributions p and q, and let ϕ_p and ϕ_q be the PDFs of the pushforward measures ${f^*}_\sharp p$ and ${f^*}_\sharp q$ respectively*[4]*. Then,*

$$\frac{\phi_p(a)}{\phi_p(a) + \phi_q(a)} = a, \qquad \forall a \in \text{supp}(\phi_p) \cup \text{supp}(\phi_q). \tag{9}$$

Intuitively this proposition states the following. Consider a point x which is mapped to a value $f^*(x) = a$. The value a characterizes the ratio of densities $p(x)/q(x)$ at x. For any other point y mapped to the same value $f^*(y) = a$, the ratio of densities $p(y)/q(y)$ is the same as $p(x)/q(x)$. Moreover, the ratio of 1D pushforward densities $\phi_p(a)/\phi_q(a)$ at a must be the same as the ratio of densities $p(x)/q(x)$ at x.

We also make a remark about Theorem 3.1.

Remark B.0.1. The result of Theorem 3.1 does not necessarily hold for other types of discriminator. For instance, the dual Wasserstein discriminator (Arjovsky et al., 2017; Gulrajani et al., 2017) does not always highlight the support difference in the original space as a support difference in the discriminator output space. This observation is formaly stated in the following proposition.

Proposition B.3. *Let $f^\star_W$ be the optimal solution of $\sup_{f:\text{Lip}(f)\leq 1} \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}[f(y)]$, where $\text{Lip}(f)$ is the Lipschitz constant of f. There exist distributions p and q such that $\text{supp}(p) \neq \text{supp}(q)$ but $\text{supp}(f^\star_{W\,\sharp} p) = \text{supp}(f^\star_{W\,\sharp} q)$.*

C. Proofs of the theoretical results

C.1. Proof of Proposition 2.1

1. $\mathcal{D}_\triangle(p, q) \geq 0$ for all $p, q \in \mathcal{P}$:

   Since $d(\cdot, \cdot) \geq 0$, for all p, q,

   $$\text{SD}(p, q) = \mathbb{E}_{x\sim p}[d(x, \text{supp}(q))] = \mathbb{E}_{x\sim p}\left[\inf_{y\sim \text{supp}(q)} d(x, y)\right] \geq 0,$$

   which makes

   $$\mathcal{D}_\triangle(p, q) = SD(p, q) + SD(q, p) \geq 0.$$

2. $\mathcal{D}_\triangle(p, q) = 0$ if and only if $\text{supp}(p) = \text{supp}(q)$:

   With statement 1, $\mathcal{D}_\triangle(p, q) = 0$ if and only if $SD(p, q) = 0$ and $SD(q, p) = 0$.

   Then,

   $$\begin{gathered} SD(p, q) = 0 \\ \Downarrow \\ \mathbb{E}_{x\sim p}[d(x, \text{supp}(q))] = 0 \\ \Downarrow \\ p\left(\{x | d(x, \text{supp}(q)) > 0\}\right) = 0. \end{gathered}$$

   This is equivalent to

   $$\forall x \in \text{supp}(p), d(x, \text{supp}(q)) = 0.$$

   Thus, $\text{supp}(p) \subseteq \text{supp}(q)$, and similarly, $\text{supp}(q) \subseteq \text{supp}(p)$,

   which makes

   $$\text{supp}(p) = \text{supp}(q).$$

[4] In order to simplify the argument, we assume that all distributions p, q, ${f^*}_\sharp p$, ${f^*}_\sharp q$ have PDFs.

C.2. Proof of Proposition 2.2

From (2), we have

$$\mathcal{D}_W^{\infty}(p,q) \coloneqq \lim_{\beta\to\infty} \mathcal{D}_W^{\beta}(p,q) = \lim_{\beta\to\infty} \inf_{\gamma\in\Gamma_\beta(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)],$$

where $\lim_{\beta\to\infty}\Gamma_\beta(p,q)$ is the set of all measures γ on $\mathcal{X}\times\mathcal{X}$ such that $\int\gamma(x,y)dy = p(x), \forall x$ and $\int\gamma(x,y)dx \leq \lim_{\beta\to\infty}(1+\beta)q(y), \forall y$.

The set of inequalities

$$\int \gamma(x,y)dx \leq \lim_{\beta\to\infty}(1+\beta)q(y), \qquad \forall y$$

can be simplified to

$$\int \gamma(x,y)dx = 0, \qquad \forall y \text{ such that } q(y)=0.$$

To put it together, we have

$$\mathcal{D}_W^{\infty}(p,q) = \inf_{\gamma\in\Gamma'(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)],$$

where $\Gamma'(p,q)$ is the set of all measures γ on $\mathcal{X}\times\mathcal{X}$ such that $\int\gamma(x,y)dy = p(x), \forall x$ and $\int\gamma(x,y)dx = 0, \forall y$ such that $q(y)=0$.

Note that

$$\mathbb{E}_{(x,y)\sim\gamma}[d(x,y)] = \int \gamma(x,y)d(x,y)dxdy.$$

Then, we can minimize $\mathbb{E}_{(x,y)\sim\gamma}[d(x,y)]$ pointwise w.r.t. x under $\gamma\in\Gamma'(p,q)$.

With any fixed x, we want to minimize

$$\int q_x(y)d(x,y)dy$$

where $\int q_x(y)dy = p(x)$ and $q_x(y)=0$ when $q(y)=0$, which implies $q_x(y)=0$ if $y\notin\operatorname{supp}(q)$.

Then clearly $q_x^*(y)$ which is equal to $p(x)$ for some y such that $d(x,y) = \inf_{y\in\operatorname{supp}(q)} d(x,y)$ and 0 otherwise is a minimizer.

Then, we arrive at

$$\inf_{\gamma\in\Gamma'(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x,y)] = \mathbb{E}_{(x,y)\sim\gamma^*}[d(x,y)],$$

where $\gamma^*(x,y) = p(x)$, for all $x\in\operatorname{supp}(p)$ and some y such that $(x,y) = \inf_{y\in\operatorname{supp}(q)} d(x,y)$. Otherwise, $\gamma^*(x,y)=0$.

Thus,

$$\mathbb{E}_{(x,y)\sim\gamma^*}[d(x,y)] = \mathbb{E}_{x\sim p}\left[\inf_{y\sim\operatorname{supp}(q)} d(x,y)\right].$$

The last equation implies $\mathcal{D}_W^{\infty}(p,q) = \mathrm{SD}(p,q)$.

Then obviously, $\mathcal{D}_W^{\infty,\infty}(p,q) \coloneqq \lim_{\beta_1,\beta_2\to\infty}\mathcal{D}_W^{\beta_1,\beta_2}(p,q) = \mathcal{D}_\triangle(p,q)$.

C.3. Proof of Proposition 2.3

1. $\mathcal{D}_W(p,q)=0$ implies $p=q$, which is equivalent to

$$\frac{p(x)}{q(x)} = 1, \qquad \forall x\in\operatorname{supp}(p)\cup\operatorname{supp}(q).$$

   Then clearly, for all finite $\beta_1,\beta_2>0$ it satisfies

$$\frac{1}{1+\beta_2} \leq \frac{p(x)}{q(x)} \leq 1+\beta_1, \qquad \forall x\in\operatorname{supp}(p)\cup\operatorname{supp}(q). \tag{10}$$

   Thus, $\mathcal{D}_W^{\beta_1,\beta_2}(p,q)=0$ for all finite $\beta_1,\beta_2>0$.

2. $\mathcal{D}_W^{\beta_1,\beta_2}(p,q) = 0$ for some finite $\beta_1, \beta_2 > 0$ is equivalent to for some finite $\beta_1, \beta_2 > 0$, (10) is satisfied.

   This implies that $\forall x \in \text{supp}(p)$, $x \in \text{supp}(q)$ and $\forall x \in \text{supp}(q)$, $x \in \text{supp}(p)$, which makes $\text{supp}(p) = \text{supp}(q)$. Thus, $\mathcal{D}_\triangle(p,q) = 0$.

3. The converse of statements 1 and 2 are false:

   (a) For all finite $\beta_1, \beta_2 > 0$, let $\text{supp}(p) = \text{supp}(q) = \{x_1, x_2\}, x_1 \neq x_2$. Let $p(x_1) = p(x_2) = 1/2$ and $q(x_1) = (1+\beta')/2$ and $q(x_2) = (1-\beta')/2$ where

   $$\beta' = \min\left(\beta_2, 1 - \frac{1}{1+\beta_1}\right).$$

   Then, it can be easily checked that (10) is satisfied, which makes $\mathcal{D}_W^{\beta_1,\beta_2}(p,q) = 0$. However, since $\beta' \neq 0$, $p \neq q$ and thus $\mathcal{D}_W(p,q) \neq 0$.

   (b) Similar to (a), let $\text{supp}(p) = \text{supp}(q) = \{x_1, x_2\}, x_1 \neq x_2$. Let $p_\varepsilon(x_1) = q_\varepsilon(x_2) = \varepsilon$ and $p_\varepsilon(x_2) = q_\varepsilon(x_1) = 1-\varepsilon$ for some $\varepsilon > 0$. Since $\text{supp}(p_\varepsilon) = \text{supp}(q_\varepsilon)$, $\mathcal{D}_\triangle(p_\varepsilon, q_\varepsilon) = 0$. However, for any finite $\beta > 0$,

   $$\lim_{\varepsilon \downarrow 0} \frac{p_\varepsilon(x_1)}{q_\varepsilon(x_1)} = \lim_{\varepsilon \downarrow 0} \frac{\varepsilon}{1-\varepsilon} = 0 < \frac{1}{1+\beta}.$$

   Thus, for any finite β_1, β_2, exists $\varepsilon > 0$ such that $\mathcal{D}_W^{\beta_1,\beta_2}(p_\varepsilon, q_\varepsilon) \neq 0$.

C.4. Proof of Proposition B.1

Consider a 2-dimensional Euclidian space $\mathbb{R}^2$ and let $\text{supp}(p) = \{(x,y) \mid x^2 + y^2 \leq 2\}$ and $\text{supp}(q) = \{(x,y) \mid 1 \leq x^2 + y^2 \leq 2\}$. Then, $\forall f^\theta(x) = \langle \theta, x \rangle$ with $\theta \in \mathbb{S}^1$,

$$\text{supp}({f^\theta}_\sharp p) = \text{supp}({f^\theta}_\sharp q) = [-2, 2].$$

C.5. Proof of Proposition B.2

For any point $a \in \text{supp}(\phi_p) \cup \text{supp}(\phi_q)$, the values of the PDFs ϕ_p and ϕ_q can be expressed as

$$\begin{aligned} \phi_p(a) &= \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}_p\left(\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}\right)}{2\varepsilon} = \lim_{\varepsilon \downarrow 0} \frac{\int_{\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}} p(x)\,dx}{2\varepsilon}, \\ \phi_q(a) &= \lim_{\varepsilon \downarrow 0} \frac{\mathbb{P}_q\left(\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}\right)}{2\varepsilon} = \lim_{\varepsilon \downarrow 0} \frac{\int_{\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}} q(x)\,dx}{2\varepsilon} \end{aligned}.$$

Note that for all $x : a - \varepsilon < f^*(x) < a + \varepsilon$ we have

$$a - \varepsilon < \frac{p(x)}{p(x) + q(x)} < a + \varepsilon,$$

which implies

$$(a-\varepsilon)(p(x)+q(x)) < p(x) < (a+\varepsilon)(p(x)+q(x)).$$

Since these inequalities hold for all $x : a - \varepsilon < f^*(x) < a + \varepsilon$, the similar relationship holds for the integrals:

$$\begin{aligned} (a-\varepsilon) \int_{\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}} (p(x)+q(x))\,dx \\ < \int_{\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}} p(x)\,dx < \\ (a+\varepsilon) \int_{\{x \mid a-\varepsilon < f^*(x) < a+\varepsilon\}} (p(x)+q(x))\,dx. \end{aligned}$$

The ratio $\phi_p(a)/(\phi_p(a)+\phi_q(a))$ can be expressed as

$$\frac{\phi_p(a)}{\phi_p(a)+\phi_q(a)} = \lim_{\varepsilon\downarrow 0} \frac{\int_{\{x|a-\varepsilon<f^*(x)<a+\varepsilon\}} p(x)\,dx}{\int_{\{x|a-\varepsilon<f^*(x)<a+\varepsilon\}} (p(x)+q(x))\,dx}.$$

Using the inequality above we observe that

$$a-\varepsilon < \frac{\int_{\{x|a-\varepsilon<f^*(x)<a+\varepsilon\}} p(x)\,dx}{\int_{\{x|a-\varepsilon<f^*(x)<a+\varepsilon\}} (p(x)+q(x))\,dx} < a+\varepsilon,$$

for all $\varepsilon > 0$, and by taking the limit $\varepsilon \downarrow 0$ we obtain

$$\frac{\phi_p(a)}{\phi_p(a)+\phi_q(a)} = a.$$

C.6. Proof of Theorem 3.1

Note that by (4) and (9), we have

$$f^*(x) = \frac{p(x)}{p(x)+q(x)} = a = \frac{\phi_p(a)}{\phi_p(a)+\phi_q(a)}, \qquad \forall x \in \operatorname{supp}(p)\cup\operatorname{supp}(q).$$

1. $\Longrightarrow$: if $\mathcal{D}_W(p,q)=0$ then $p=q$, then ${f^*}_\sharp p = {f^*}_\sharp q$. Thus, $\mathcal{D}_W({f^*}_\sharp p, {f^*}_\sharp q)=0$.
   $\Longleftarrow$: if $\mathcal{D}_W({f^*}_\sharp p, {f^*}_\sharp q)=0$, then ${f^*}_\sharp p = {f^*}_\sharp q$, which implies $\phi_p(a)=\phi_q(a), \forall a$. Then, we have
   $$p(x)=q(x), \qquad \forall x \in \operatorname{supp}(p)\cup\operatorname{supp}(q).$$
   Thus, $p=q$ and $\mathcal{D}_W(p,q)=0$.
2. $\Longrightarrow$: if $\mathcal{D}_W^{\beta_1,\beta_2}(p,q)=0$, (10) holds, then
   $$\frac{1}{2+\beta_2} \le f^*(x) \le \frac{1+\beta_1}{2+\beta_1}, \qquad \forall x \in \operatorname{supp}(p)\cup\operatorname{supp}(q).$$
   This implies that
   $$\frac{1}{2+\beta_2} \le \frac{\phi_p(a)}{\phi_p(a)+\phi_q(a)} \le \frac{1+\beta_1}{2+\beta_1}, \qquad \forall a \in \operatorname{supp}(\phi_p)\cup\operatorname{supp}(\phi_q).$$
   Thus, we have
   $$\frac{1}{1+\beta_2} \le \frac{\phi_p(a)}{\phi_q(a)} \le 1+\beta_1, \qquad \forall a \in \operatorname{supp}(\phi_p)\cup\operatorname{supp}(\phi_q).$$
   This leads to $\mathcal{D}_W^{\beta_1,\beta_2}({f^*}_\sharp p, {f^*}_\sharp q)=0$.
   $\Longleftarrow$: similarly, when $\mathcal{D}_W^{\beta_1,\beta_2}({f^*}_\sharp p, {f^*}_\sharp q)=0$,
   $$\frac{1}{1+\beta_2} \le \frac{\phi_p(a)}{\phi_q(a)} \le 1+\beta_1, \qquad \forall a \in \operatorname{supp}(\phi_p)\cup\operatorname{supp}(\phi_q)$$
   $$\Downarrow$$
   $$\frac{1}{2+\beta_2} \le f^*(x) \le \frac{1+\beta_1}{2+\beta_1}, \qquad \forall x \in \operatorname{supp}(p)\cup\operatorname{supp}(q)$$
   Thus, (10) holds and $\mathcal{D}_W^{\beta_1,\beta_2}(p,q)=0$.
3. Note that $f^*(x)=1$ if and only if $p(x)>0$, $q(x)=0$, and $f^*(x)=0$ if and only if $q(x)>0$, $p(x)=0$.
   We also have $\phi_p(0)>0$ iff $0\in\operatorname{supp}(\phi_p)$ and $\phi_q(1)>0$ iff $1\in\operatorname{supp}(\phi_q)$.
   Thus, $f^*(x)=1$ for some x if and only if $\operatorname{supp}(p)\setminus\operatorname{supp}(q)\neq\varnothing$ and $x\in\operatorname{supp}(p)\setminus\operatorname{supp}(q)$, in which case $\phi_p(1)>0$, which makes $\operatorname{supp}(\phi_p)\setminus\operatorname{supp}(\phi_q)\neq\varnothing$.
   Similarly, $f^*(x)=0$ for some x if and only if $\operatorname{supp}(q)\setminus\operatorname{supp}(p)\neq\varnothing$ and $x\in\operatorname{supp}(q)\setminus\operatorname{supp}(p)$, in which case $\phi_q(0)>0$, which makes $\operatorname{supp}(\phi_q)\setminus\operatorname{supp}(\phi_p)\neq\varnothing$.
   Then, we have $\operatorname{supp}(p)\neq\operatorname{supp}(q)$ if and only if $\operatorname{supp}(\phi_p)\neq\operatorname{supp}(\phi_q)$.
   This is equivalent to $\mathcal{D}_\triangle(p,q)=0$ if and only if $\mathcal{D}_\triangle({f^*}_\sharp p, {f^*}_\sharp q)=0$.

C.7. Proof of Proposition B.3

Consider a 1-dimensional Euclidian space $\mathbb{R}$. Let $\mathrm{supp}(p) = \{0, 1\}$ with $p(0) = 3/4$ and $p(1) = 1/4$. Let $\mathrm{supp}(q) = \{-1, 0, 1\}$ with $q(-1) = 1/4$, $q(0) = 1/4$ and $q(1) = 1/2$.

Clearly, $\mathrm{supp}(p) \neq \mathrm{supp}(q)$.

Let $f^\star_W$ be the optimal solution of

$$\sup_{f:\mathrm{Lip}(f)\leq 1} \mathbb{E}_{x\sim p}[f(x)] - \mathbb{E}_{y\sim q}[f(y)].$$

Then, the value of $f^\star_W$ only matters at $\{-1, 0, 1\}$. So we have $f^\star_W$ be the optimal solution of

$$\sup_{f:\mathrm{Lip}(f)\leq 1} \frac{-f^\star_W(-1) + 2f^\star_W(0) - f^\star_W(1)}{4}.$$

By symmetry, $f^\star_W(-1) = f^\star_W(1)$, making $\mathrm{supp}(f^\star_{W\sharp}p) = \mathrm{supp}(f^\star_{W\sharp}q)$.

D. Discussion of "soft" and "hard" assignments for OT with 1D discrete distributions

Direct application of the OT problem (2) to 1D discrete distributions $\widehat{p}_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(x)$, $\widehat{q}_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}(x)$ gives

$$\widehat{\mathcal{D}}^c_W(\widehat{p}_n, \widehat{q}_n) = \min_{\widehat{\gamma}\in\widehat{\Gamma}_c(\widehat{p}_n,\widehat{q}_n)} \sum_{i=1}^n\sum_{j=1}^n \widehat{\gamma}_{ij} d(x_i, y_j), \tag{11}$$

where $c > 0$ and $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ is the set of probability measures $\widehat{\gamma}$ on $\overline{1,n} \times \overline{1,n}$ such that $\sum_{j=1}^n \widehat{\gamma}_{ij} = \frac{1}{n}, \forall\, i \in \overline{1,n}$ and $\sum_{i=1}^n \widehat{\gamma}_{ij} \leq \frac{c}{n}, \forall\, j \in \overline{1,n}$.

The optimization goal in (11) is to find a "soft-assignment" $\widehat{\gamma}$ which describes the transportation of probability mass from points x_i in $\widehat{p}_n$ to points y_j in $\widehat{q}_n$. The parameter c controls the set of admissible assignments $\widehat{\Gamma}_c$. The constraints imposed on the assignments $\widehat{\gamma}$ are similar to the constraints on β-admissible coupling $\gamma \in \Gamma_\beta$ in (2): under a c-admissible assignment $\widehat{\gamma}$, the total mass of points in $\widehat{p}_n$ transported to each of the points y_j can not exceed $\frac{c}{n}$. Note that if one additionally requires that the transportation plan $\widehat{\gamma}$ implements a "hard-assignment" ($\widehat{\gamma}_{ij} \in \{0, \frac{1}{n}\}$) which assigns each point x_i to exactly one point y_j, then c gives the upper-bound on the number of points x_i which can be transported to any given point y_j. In fact, it can be shown (see discussion below) that for integer values of $c = 1, 2, \ldots$ the set of minimizers of (11) must contain a "hard-assignment" transportation plan. Motivated by this observation, we can reduce the original "soft-assignment" problem (11) to the "hard-assignment" problem (6).

We claim that in "soft-assignment" OT problem (11) with integer c the set of minimizers of (11) must contain a "hard-assignment" transportation plan $\widehat{\gamma} : \widehat{\gamma}_{ij} \in \{0, \frac{1}{n}\}$. Below we justify this claim.

Note that for $c = 1$ the OT problem (11) is the standard OT problem for Wasserstein distance, since the inequality constraints $\sum_{i=1}^n \widehat{\gamma}_{ij} \leq \frac{1}{n}, \forall\, j \in \overline{1,n}$ can only be satisfied as equalities. For this problem, it is known (e.g. see (Peyré et al., 2019) Proposition 2.1) that the set of optimal couplings contains a "hard-assignment" represented by a normalized permutation matrix. This fact can be proven using the Birkhoff–von Neumann theorem. The Birkhoff-von Neumann theorem states that the set of doubly stochastic matrices

$$P \in \mathbb{R}^{n\times n}: \qquad P_{ij} \geq 0, \forall\, i, j \in \overline{1,n}, \qquad \sum_{j=1}^n P_{ij} = 1, \forall\, i \in \overline{1,n}, \qquad \sum_{i=1}^n P_{ij} = 1, \forall\, j \in \overline{1,n},$$

is exactly the set of all finite convex combinations of permutation matrices. In the context of the linear program (11) with $c = 1$, the Birkhoff-von Neumann means that all extremal points of the polyhedron $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ are hard-assignment matrices. Therefore, by the fundamental theorem of linear programming (Bertsimas & Tsitsiklis, 1997), the minimum of the objective is reached at a "hard-assignment" matrix $\widehat{\gamma}$.

We argue that a similar result holds for the case of integer $c > 1$. In this case, the matrices in $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ can not be associated with the doubly stochastic matrices, since constraints on of the marginals of $\widehat{\gamma}$ are relaxed to inequality constraints. Because

of that, the Birkhoff-von Neumann theorem can not be applied. However, Budish et al. (2009) provide a generalization of the Birkhoff-von Neumann theorem (Theorem 1 in (Budish et al., 2009)) which applies to the cases where the equality constraints are replaced with integer-valued inequality constraints (recall that we consider integer c). Using this generalized result, our claim can be proven in the following sequence of steps.

Clearly, the polyhedron $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ contains all "hard-assignment" matrices and all their finite convex combinations. The result proven in (Budish et al., 2009) implies that each element of $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ can be represented as a finite convex combination of "hard-assignment" matrices. Thus, the polyhedron $\widehat{\Gamma}_c(\widehat{p}_n, \widehat{q}_n)$ is exactly the set of all finite convex combinations of "hard-assignment" matrices and all extremal points of the polyhedron are "hard-assignment" matrices. Finally, by analogy with the case of $c = 1$, we invoke the fundamental theorem of the linear programming and conclude that the minimum of the objective (11) is reached at $\widehat{\gamma}$ corresponding to "hard-assignment" matrix.

E. Details of the ASA algorithm

This section provides details of the "mini-batch + history" update of the ASA algorithm mentioned in Section 4. We create two history buffers: $h^p = \{h_i^p\}_{i=1}^N$, storing the discriminator outputs h_i^p for for the last N samples from p_θ, and the similar buffer $h^q = \{h_i^q\}_{i=1}^N$ for q_θ. For a new pair of mini-batches $\{x_i\}_{i=1}^n \sim p_\theta$ and $\{y_i\}_{i=1}^n \sim q_\theta$, instead of the mini-batch assignments used in (7), we find "mini-batch + history" assignments $\pi^h_{p\to q}$ and $\pi^h_{q\to p}$ between the combined "mini-batch + history" point sets $v^p = (g_\phi(x_1), \dots g_\phi(x_n), h_1^p, \dots, h_N^p)$ and $v^q = (g_\phi(y_1), \dots g_\phi(y_n), h_1^q, \dots, h_N^q)$. The complete specification of the ASA training procedure is provided in Algorithm E.1.

E.1. History size experiment

In Section 4 we mentioned the effects of mini-batch training for our proposed method. To quantify this effect, we explore ASA with different size of history buffers under the same setting described in Section 5 with the label distribution shift corresponding to $\alpha = 1.5$. When history size is 0, it means that we are training with current mini-batch only (eq. (7)).

In Table E.1 we report the results of the evaluation of ASA with different history sizes. As in MNIST-to-USPS experiment above, we compute the class average accuracy on target data. In order to quantify the alignment achieved by the methods, we also report support and distribution distances between the pushforward distributions induced by the discriminator on source and target data. We evaluate both support and distribution distance on the entire training populations rather than batches. The support distance is the mean squared transportation distance under the optimal ∞-to-1 assignment described in Section 3.2. The distribution distance is computed as the mean square transportation distance under the optimal 1-to-1 assignment described in Section 3.2. For both support and distribution distances, lower values correspond to better alignment of supports/distributions.

From Table E.1, comparing with the source only baseline, alignment methods are enforcing their desired alignments and these constraints are not satisfied with source only training. Comparing with DANN, we can see that with smaller size of history, ASA performs more like distribution alignment, while all history sizes are enough for support alignment. Also, note the correlation between distribution distance and target accuracy: with a shift in label distribution, when distribution alignment is achieved, the target accuracy suffers.

Table E.1. Analysis of effect history size parameter for ASA on MNIST-to-USPS data with different the class label distribution shift corresponding to $\alpha = 1.5$. We report the mean and 1 standard deviation values across 3 runs.

Method	History size	Target acc (%)	Support distance	Distribution distance
Source only	—	14.24 ± 3.34	95.22 ± 45.63	310.55 ± 131.22
DANN	—	48.39 ± 0.25	0.000003 ± 0.000001	0.00074 ± 0.00008
ASA-batch	0	50.39 ± 1.41	0.000025 ± 0.000026	0.016 ± 0.002
ASA	100	68.99 ± 11.97	0.000019 ± 0.000013	0.472 ± 0.239
ASA	1000	77.16 ± 6.01	0.000175 ± 0.000068	3.518 ± 0.682
ASA	3000	72.72 ± 6.89	0.0010 ± 0.0011	6.111 ± 0.195

Algorithm E.1 Adversarial support alignment (ASA)

Input:

- p_θ, q_θ: distributions with parameters θ;
- θ_0: initial parameters of p, q;
- g_ϕ: discriminator network with parameters ϕ;
- ϕ_0: initial parameters of g;
- n: mini-batch size;
- N: maximum history buffer size;
- T: number of training steps for θ;
- K: number of g_ϕ updates per one update of p_θ, q_θ;
- d: distance on the 1D discriminator output space.

def UPDATEHISTORY$(h, \{a_i\}_{i=1}^n)$
    Append values $\{a_i\}_{i=1}^n$ to the end of history buffer h.
    if $\text{size}(h) > N$: remove the oldest $N - \text{size}(h)$ elements from h.

$\theta \leftarrow \theta_0,\ \phi \leftarrow \phi_0$.
$h^p \leftarrow [\,],\ h^q \leftarrow [\,]$.
for $t = 1, \ldots, T$
    for $k = 1, \ldots K$
        Sample mini-batches $\{x_i\}_{i=1}^n \sim p_\theta$, $\{y_i\}_{i=1}^n \sim q_\theta$.
        UPDATEHISTORY$(h^p, \{g_\phi(x_i)\}_{i=1}^n)$, UPDATEHISTORY$(h^q, \{g_\phi(y_i)\}_{i=1}^n)$.
        Perform optimization step on ϕ using stochastic gradient

$$\nabla_\phi \left(\frac{1}{n} \sum_{i=1}^{n} \Big[\log(1 + \exp(-g_\phi(x_i))) + \log(1 + \exp(g_\phi(y_i))) \Big] \right).$$

    Sample mini-batches $\{x_i\}_{i=1}^n \sim p_\theta$, $\{y_i\}_{i=1}^n \sim q_\theta$.
    $v^p \leftarrow \text{concat}([g_\phi(x_i)]_{i=1}^n, h^p),\ v^q \leftarrow \text{concat}([g_\phi(y_i)]_{i=1}^n, h^q)$.
    Compute optimal ∞-to-1 assignment between $\{g_\phi(x_i)\}_{i=1}^n$ and v^q:

$$\pi^h_{p\to q}(i) = \operatorname*{argmin}_j d\left(g_\phi(x_i), v^q_j\right).$$

    Compute optimal ∞-to-1 assignment between $\{g_\phi(y_i)\}_{i=1}^n$ and v^p:

$$\pi^h_{q\to p}(i) = \operatorname*{argmin}_j d\left(g_\phi(y_i), v^p_j\right).$$

    Perform optimization step on θ using stochastic gradient

$$\nabla_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \Big[d\left(g_\phi(x_i), v^q_{\pi^h_{p\to q}(i)}\right) + d\left(g_\phi(y_i), v^p_{\pi^h_{q\to p}(i)}\right) \Big] \right).$$

    */ ⋆ *The gradient of this expression w.r.t. θ is propagated*
    */ ⋆ *through the current mini-batch examples $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$.*
    */ ⋆ *The values stored in the history buffers h^p, h^q do not depend on θ.*
    UPDATEHISTORY$(h^p, \{g_\phi(x_i)\}_{i=1}^n)$, UPDATEHISTORY$(h^q, \{g_\phi(y_i)\}_{i=1}^n)$.

F. MNIST-to-USPS results

Table F.1. Class average accuracy (%) on target domain of the different adaptation methods on MNIST-to-USPS data with different levels of class label distribution shift. We report the mean and 1 standard deviation values across 5 runs.

Method	$\alpha = 0$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$
Source only	27.71 ± 11.11	22.52 ± 5.69	15.85 ± 3.26	15.46 ± 5.80
DANN	91.61 ± 0.90	71.13 ± 1.55	47.75 ± 0.88	31.32 ± 1.96
DANN-OT	97.82 ± 0.16	94.85 ± 1.60	56.57 ± 3.25	29.96 ± 2.80
sDANN-2	95.77 ± 3.79	95.67 ± 3.74	64.33 ± 8.06	26.35 ± 7.82
sDANN-4	91.81 ± 4.53	91.46 ± 4.51	77.35 ± 12.94	41.96 ± 8.16
sDANN-16	83.21 ± 5.84	80.74 ± 3.39	73.66 ± 13.13	39.88 ± 23.75
ASA (ours)	93.57 ± 1.10	88.49 ± 5.79	80.44 ± 6.20	48.11 ± 8.63