A Unified Framework for Comparing Distribution Matching Methods Across Trustworthy Machine Learning Tasks

Brian Ko[1] **Ziyu Gong**[1] **Jim Lim**[1] **David I. Inouye**[1]

[1]Elmore Family Dept. of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Abstract

Distribution matching (DM) is a fundamental tool in trustworthy machine learning (TML), with applications in fairness, calibration, and domain adaptation. While prior work advances individual DM methods based on information-theoretic and geometric divergences, a unified comparative framework remains lacking. We propose a framework integrating DM methods, metrics, and TML tasks to enable systematic comparisons. To our knowledge, this is the first work to compare latent spaces in TML while addressing scaling inconsistencies via ZCA whitening. We empirically evaluate MMD, Sinkhorn, adversarial, and VAE-based DM methods across fairness, calibration, and domain adaptation. Our findings reveal: (1) accuracy and expected calibration error (ECE) are positively correlated in low-accuracy models but negatively correlated in high-accuracy models, extending prior results Tao et al. [2023a]; (2) logit-based fairness methods outperform latent-based approaches; and (3) strict DM enforcement can reduce target accuracy in domain adaptation, challenging theoretical bounds Ben-David et al. [2006b]. These insights inform the selection and refinement of DM algorithms for TML applications.

1 INTRODUCTION

Domain-invariant representation learning (DIRL) Zhao et al. [2019, 2022] aims to learn a representation function $g_\theta : \mathbb{X} \to \mathbb{Z}$, which map data from different domains into a shared latent space where their distributions align, enabling models to focus on task-relevant features while ignoring domain-specific variation as shown in Figure 1. Unlike representation learning for classification which seeeks to maximize the divergence between class distributions, DIRL seeks to minimize the divergence between the domain distribution. Hence, DIRL can be seen as the natural complement to classification by defining what is not important, while classification defines what is important. This approach is foundational to many trustworthy machine learning (TML) tasks, such as fair classification (invariance to sensitive attributes), domain adaptation (aligning source and target environments), and uncertainty calibration (matching prediction confidence across subgroups). By minimizing distributional divergence in the latent space, DIRL addresses the pervasive challenge of distribution shift, which violates the standard independent and identically distributed (IID) assumption and undermines model reliability in real-world applications.

Prior work on distribution matching has primarily been developed within specific TML tasks, often focusing on individual approaches rather than a comparative or unified framework Han et al. [2023b], Reddy [2022b], Tao et al. [2023b], Gulrajani and Lopez-Paz [2020], Marx et al. [2024b]. For instance, in uncertainty calibration Marx et al. [2024b], DM has been explored using kernel-based approaches such as Maximum Mean Discrepancy (MMD) to align predicted and true confidence distributions. In contrast, domain adaptation methods typically rely on adversarial learning, where generative adversarial networks (GANs) or domain classifiers enforce domain-invariant representations Ganin et al. [2016b]. In fairness, logit-based methods enforce fairness by directly constraining output distributions Chung et al. [2024a], while latent space-based methods align intermediate feature distributions Madras et al. [2018]. Despite the diversity of DM techniques, they are often developed in isolation, without a comprehensive comparison across TML applications. Consequently, there is limited understanding of which DM methods generalize best across tasks or how different alignment techniques trade off between computational efficiency, stability, and effectiveness.

To bridge this gap, we propose a unified framework for systematically comparing DM methods across multiple TML tasks. Our framework integrates representative DM methods including Maximum Mean Discrepancy (MMD) Gretton

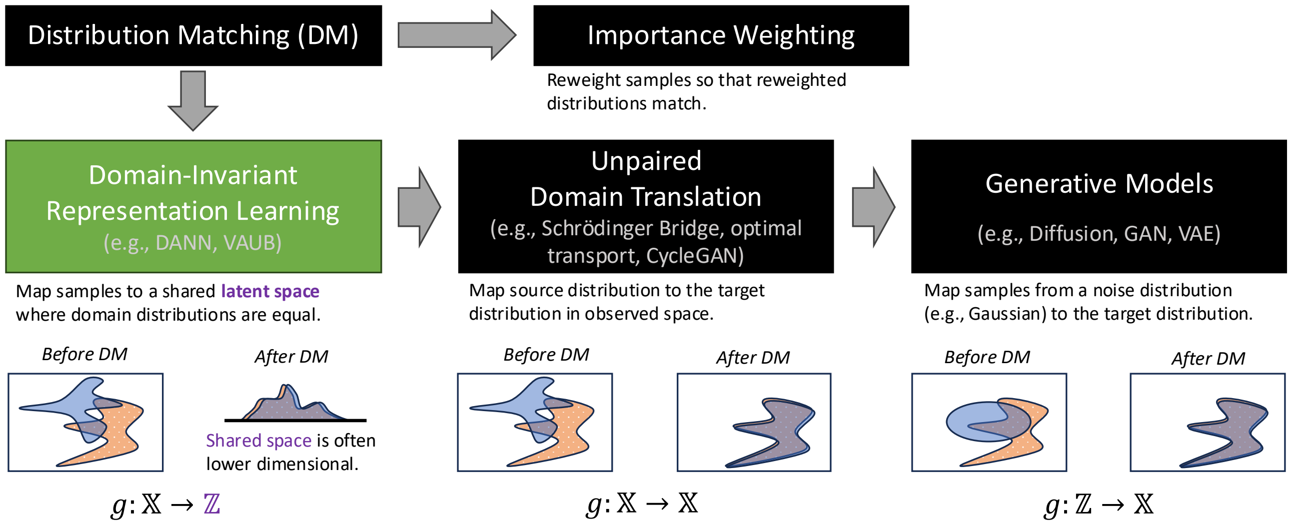

Figure 1: Distribution matching (DM) aims to map two or more distributions to the same distribution. The most general form is (unpaired) domain-invariant representation learning (left) where the algorithm only has access to samples from each domain but can project them into a (lower-dimensional) latent space. Unsupervised domain translation is a special case in which one distribution is the source and one is the target, which does not change. Finally, an even more special case is generative models which maps from a known distribution (usually Gaussian) to the data distribution. Importance weighting is another technique for DM but is not considered in this work because most trustworthy ML applications require a new representation rather than sample weights.

et al. [2012], Sinkhorn divergence Feydy et al. [2019a], adversarial domain alignment Ganin et al. [2016b], and VAE-based methods Gong et al. [2024], and evaluates their performance on three major TML tasks: fairness, calibration, and domain adaptation. Unlike prior work that considers DM in isolation for a single task, our framework enables direct comparison across these tasks, providing insights into the relationship between intrinsic metric (MMD, Sinkhorn) with task specific metric (ECE Guo et al. [2017b], DP Han et al. [2023b]). Additionally, we introduce a normalized divergence metric to control for latent space scaling, ensuring fair evaluation DM methods using non-parametric geometric divergences that can be computed directly from samples.

Our empirical results reveal key trends in DM effectiveness across tasks and expose limitations in current methods that future research must address. First, recent study on calibration revealed that there is negative correlation within accuracy and expected calibration error (ECE) on strong predictive models, but our study finds that there is positive correlation between accuracy and ECE on low accuracy models. Moreove, strong distribution matching can improve accuracy but may induce overconfidence, highlighting the importance of post-hoc adjustment. Second, current research on fairness mostly focus on representation learning (latent based) method, but we find that logit based method outperforms the latent based method. Lastly, our empirical findings contradict widely used theoretical bounds grounded in the $\mathcal{A}-$distance Ben-David et al. [2006b] that there should be gap between source and target distribution in order to achieve best performance. Through this benchmark, we aim to guide the selection of DM techniques for TML applications and inspire the development of more robust, generalizable DM algorithms. Our contribution can be summarized as follows:

1. We formalize a common theoretical framework that integrates DIRL and DM methods under a single umbrella, enabling systematic comparisons.
2. We provide theoretical connection between DM methods with information theocratic divergence Pardo and Vajda [2003] and geometric divergence Amari [2009] to provide strength and weakness of each DM methods.
3. Using the unified DM framework, we evaluate different DM methods across fairness, domain adaptation, and calibration tasks, highlighting their connection between intrinsic metric (MMD, Sinkhorn) and task specific metric (DP, ECE), and provide insightful guideline for practical usage.

2 UNIFIED FRAMEWORK FOR DISTRIBUTION MATCHING AND TRUSTWORTHY ML TASKS.

Notation. Let $\mathrm{x} \in \mathbb{X}$, $\mathrm{y} \in \mathbb{Y}$, and $\mathrm{d} \in \{1, 2, \ldots, k\}$ denote random variables corresponding to the input, target (optional), and domain label, respectively. Let $\mathbf{z} := g_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon) \sim p_\theta(\mathbf{z}|\mathrm{y}, \mathbf{x}, \mathrm{d})$ denote the latent representation of $\mathbf{x}$, and for logit based method, we denote $\hat{\mathbf{q}} := g_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon) \sim p_\theta(\hat{\mathbf{q}}|\mathrm{y}, \mathbf{x}, \mathrm{d})$ where g_θ is called the *matcher* with parameters θ that may optionally depend on the target variable y, the domain d, and exogenous noise ϵ to encompass stochastic aligners. If g does not depend on d and/or ϵ we will suppress notation w.r.t. these random variables for notational simplicity. Let $p_{\mathrm{data}}(\mathbf{x}, \mathrm{y}, \mathrm{d})$ denote the true data distribution. Let ϕ denote parameters of variational models or distributions, e.g., $q_\phi(\mathbf{x}, \mathrm{y}, \mathrm{d})$ will denote a variational distribution and $h_\phi(\mathbf{z})$ will denote a variational discriminator for adversarial learning. Let ψ denote application-specific parameters, e.g., $\hat{\mathrm{y}} := f_\psi(\mathbf{z})$ will denote the predicted class based on the given classifier head in fair classification. Entropy, cross entropy, and mutual information will be denoted by $\mathrm{H}(\mathbf{x})$ and $\mathrm{H_c}(\mathbf{x}, \mathbf{z})$, and $\mathrm{I}(\mathbf{x}, \mathbf{z})$, respectively. Let $\mathrm{D}(p, q)$ denote a distribution divergence between p and q, e.g., $\mathrm{D_{KL}}$, $\mathrm{D_{JSD}}$, and $\mathrm{D}_{\mathrm{W}_\rho}$ will denote KL, JSD, and Wasserstein-ρ divergences, respectively. Similarly, let $\widehat{\mathrm{D}}$, $\overline{\mathrm{D}}$, and $\underline{\mathrm{D}}$ denote an approximation, an upper bound, or a lower bound of a divergence respectively. Because DM involves mininizing a divergence w.r.t. the matcher parameters θ, we will let $\mathrm{D}(\theta) := \mathrm{D}(p_\theta(\mathbf{z}|\mathrm{d}{=}1), p_\theta(\mathbf{z}|\mathrm{d}{=}2))$ with slight abuse of notation.

Distribution Matching Problem. The distribution matching problems we consider can be formulated as a task-specific objective plus a distribution matching constraint on the matched representation.

Definition 1. *(Distribution Matching Problem). A distribution matching problem minimizes a task objective* $\mathrm{L_{task}}(\tilde{f}_\psi, \tilde{g}_\theta)$*, where* $\tilde{f}_\psi$ *is a task-specific model and* $\tilde{g}_\theta$ *is the* matcher *model, subject to a DM constraint on the matched representation* $\tilde{\mathbf{z}} := \tilde{g}_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon)$*:*

$$\min_{\psi,\theta} \ \mathrm{L_{task}}(\tilde{f}_\psi, \tilde{g}_\theta) \quad \textit{s.t.} \quad \mathrm{D}(p_\theta(\tilde{\mathbf{z}}|\mathrm{d}{=}1), p_\theta(\tilde{\mathbf{z}}|\mathrm{d}{=}2)) \leq \delta \tag{1}$$

where $\mathrm{D}(\cdot,\cdot)$ *is a distribution divergence and* δ *is the DM slackness hyperparameter.*

In practice, minimizing a distribution divergence is challenging given only samples. Most approaches use tractable and differentiable approximations to well-known divergences. We will first explain common loss functions that aim to solve trustworthy ML tasks and then review the main approaches to minimizing a distribution divergence.

2.1 UNIFIED FORMALIZATION OF TRUSTWORTHY ML TASKS AS DISTRIBUTION MATCHING

Many trustworthy ML tasks can be formulated as DM problems. In some cases, DM is fundamental to the trustworthy ML task (e.g., fairness or calibration), while in others, DM is one approach to the task (e.g., domain adaptation). For the tasks where DM is fundamental, the key question is: *What is the empirically achievable Pareto frontier between the task objective and the DM constraint (e.g., fairness-accuracy tradeoff)?* For the tasks where DM is an approach, the key question is: *Is DM performance correlated with the relevant task performance (e.g., does better DM yield better domain adaptation performance)?* In particular, we would like to disentangle the effect of the DM algorithm—which may be far from optimal—from the task performance. We conjecture that in some cases, the DM algorithm fails to achieve the DM objective even though the task objective may be reasonable.

Group Fair ML as Distribution Matching The goal of fair learning is to be as accurate as possible while satisfying a fairness constraint. Demographic parity (DP) (also known as statistical parity) is one common notion of group fairness that is satisfied if and only if $p(\hat{\mathrm{y}} = 1|\mathrm{d}{=}1) = p_\theta(\hat{\mathrm{y}} = 1|\mathrm{d}{=}2)$, i.e., these two distributions match. Fair classification seeks to directly learn predictions that are fair. Fair representation learning seeks to learn a representation such that all downstream tasks will be fair. We unify fair learning under our DM framework and notation below.

Proposition 1. *Fair learning Madras et al. [2018], Song et al. [2019b] w.r.t. DP is a DM problem (1) with* $\tilde{g}_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon) = g_\theta(\mathbf{x}, \epsilon)$ *and* $\mathrm{L_{task}} = \mathbb{E}[\ell(f_\psi(g_\theta(\mathbf{x}, \epsilon)), \mathrm{y})]$ *for fair classification and* $\mathrm{L_{task}} = -\mathrm{I}(\mathbf{x}, \mathbf{z} = g_\theta(\mathbf{x}, \epsilon)|\mathrm{d})$ *for fair representation learning.*

In practice, both the classification and mutual information task objectives are often combined (e.g., [Madras et al., 2018, Gong et al., 2024] approximate mutual information via a VAE objective).

Calibration as DM Problem Canonical calibration Vaicenavicius et al. [2019b] means that the predicted probabilities for all classes match the true probabilities:

$$p(\mathrm{y} = y|\hat{\mathbf{q}}) = p(\hat{\mathrm{y}} = y|\hat{\mathbf{q}}) := \mathrm{q}_y, \quad \forall y \in \mathbb{Y}, \hat{\mathbf{q}} \in \Delta^{|\mathbb{Y}|} \tag{2}$$

where $\hat{\mathbf{q}} := g_\theta(\mathbf{x})$ is the predicted class probabilities for k classes and $\Delta^{|\mathbb{Y}|}$ denotes the probability simplex. This calibration condition is a type of *conditional* distribution matching problem, i.e., match the marginal distribution of predictions to the true distribution *conditioned* on the model's output $\mathbf{q}$. In this case, the domain label is whether it is the real target variable or the predicted target variable. Marx et al. [2023] showed that indeed many types of calibration including regression, classification, and decision calibration can be framed as conditional distribution matching problems. In fact, because the marginal distribution of the conditioning variables is the same regardless of the domain, the problem can be equivalently written as a unconditional DM problem. We now unify the results from Marx et al. [2023] using our framework below.

Proposition 2. *Calibration during training is DM Marx et al. [2023] Letting* $\hat{y}'$, y', *and* c *denote the forecast, target, and conditioning variables from Marx et al. [2023, Tables 1 and 2], respectively, where* $\hat{y}'$ *and* y' *are functions of* $\hat{y}$ *and* y,[1] *calibration during training is a DM problem (1) with* $\tilde{g}_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon) = (\tilde{\mathrm{y}}', g_\theta(\mathbf{x}, \epsilon))$, *where* $\tilde{\mathrm{y}}' = \mathbb{1}(\mathrm{d}{=}1)\hat{\mathrm{y}}' + \mathbb{1}(\mathrm{d}{=}2)\mathrm{y}'$ *selects between the forecasted and target variables,* $c = g_\theta(\mathbf{x}, \epsilon)$ *represents the conditioning variable, and* $\mathrm{L}_{\text{task}} = \mathbb{E}[\ell(g_\theta(\mathbf{x}, \epsilon), \mathrm{y})]$ *is the standard negative log-likelihood ERM objective.*

Domain Adaptation via Domain-Invariant Features Inspired by the bounds on domain adaptation generalization in Ben-David et al. [2006a], many domain adaptation papers aim to learn domain-invariant features, i.e., latent features whose distribution is independent of the domain labels. Specifically, Ben-David et al. [2006a] showed that the risk on the target domain could be bounded by the risk on the source domain plus the divergence between the feature distributions and a constant. A natural approach is to reduce the divergence between the feature distributions, i.e., distribution matching. Thus, domain-invariant domain adaptation can be unified under our framework.

Proposition 3 (Domain-invariant domain adaptation is DM)**.** *Domain-invariant domain adaptation is a DM problem (1) with* $\tilde{g}_\theta(\mathbf{x}, \mathrm{y}, \mathrm{d}, \epsilon) = g_\theta(\mathbf{x}, \epsilon)$ *and* $\mathrm{L}_{\text{task}} = \mathbb{E}[\ell(f_\psi(g_\theta(\mathbf{x}, \epsilon)), \mathrm{y})]$ *is the standard ERM objective where* f_ψ *is the classification head on top of the domain-invariant representation* $\mathbf{z} = g_\theta(\mathbf{x}, \epsilon)$.

2.2 UNIFIED TRAINING OBJECTIVES FOR DISTRIBUTION MATCHING

2.2.1 Comparison between Information-Theoretic Divergence vs Geometric Divergence

We broadly categorize differentiable divergences into information-theoretic and geometric. Information-theoretic divergences Amari and Cichocki [2010] are usually estimated using a variational approximation. Information-theoretic divergences have the elegant property of being invariant under invertible transformations [Qiao and Minematsu, 2008] and thus are very useful when operating in latent spaces where the scale is irrelevant. Moreover, it can be computed with $O(N)$ for discrete measure Séjourné et al. [2023]. The drawbacks are information-theoretic divergences usually require learning an auxiliary variational model, which may be challenging itself and it is sensitive to support mismatch Séjourné et al. [2023]. Geometric divergences on the other hand use distances between points in the space and thus vary with scale Amari [2009]. This makes it more challenging to apply geometric-based divergences in latent space as simple scaling transformations drastically change these divergence measures. However, the two most common geometric divergences, Wasserstein and MMD, can be non-parametrically approximated using only a batch of samples from both domains without the need to train an auxiliary model. Also, geometric divergence metrize weak* topology that is $\alpha_n \rightharpoonup \alpha \Leftrightarrow L(\alpha_n, \alpha) \to 0$, which implies that a lower loss corresponds to closer distribution matching Feydy et al. [2019b].

2.2.2 Information Theoretic Divergences via Parametric Variational Bounds

Most differentiable approximations to information-theoretic divergences are bounds that involve training a variational model h_ϕ, such that the bound is tight if optimized perfectly but otherwise remains a bound. Adversarial GAN-based approaches form a variational *lower* bound on a divergence. The standard GAN-based loss bounds the JS divergence and trains a classifier with cross entropy loss ℓ_{CE} to predict the domain label:

$$\underline{\mathrm{D}}_{\text{ADV}}(\theta) := \max_\phi \mathbb{E}_p[-\ell_{\text{CE}}(h_\phi \circ g_\theta(\mathbf{x}), \mathrm{d})] \le \mathrm{D}_{\text{JSD}}(\theta)\,. \tag{3}$$

Adversarial objectives for all f-divergences Sason and Verdú [2016] and even Wasserstein distance Panaretos and Zemel [2019] (a geometric divergence) can be formulated. Notice that the DM problem involves minimizing this approximation and thus it forms a min-max, i.e., adversarial problem, hence the name.

In contrast to adversarial lower bounds, there have been multiple approaches to form variational *upper* bounds. One of the more common bounds is based on a variational autoencoder (VAE) structure. Recently, [Gong et al., 2024] generalized previous VAE-based approaches into a self-contained loss similar to the adversarial loss above that upper bounds the JSD:

$$\overline{\mathrm{D}}_{\text{VAUB}}(\theta) := \min_\phi \mathbb{E}_p\Big[-\log\Big(\frac{q_\phi(\mathbf{x}|\mathbf{z}, \mathrm{d})}{p_\theta(\mathbf{z}|\mathbf{x}, \mathrm{d})} \cdot q_\phi(\mathbf{z})\Big)\Big] + C \tag{4}$$

$$\ge \mathrm{D}_{\text{JSD}}(\theta)\,, \tag{5}$$

where $g_\theta(\mathbf{x}; \mathrm{d}, \epsilon)$ is a *stochastic* encoder using the reparameterization trick where $\epsilon \sim \mathcal{N}(0, I)$, $q_\phi(\mathbf{x}, \mathbf{z}|\mathrm{d}) := q_\phi(\mathbf{z})q_\phi(\mathbf{x}|\mathbf{z}, \mathrm{d})$ is a decoder distribution where $q_\phi(\mathbf{z})$ is a learnable prior distribution, and C is a constant that is

[1] Note that in many cases, $\hat{y}' = \hat{y}$ and similarly $y' = y$, but there are some cases from Marx et al. [2023] such as quantile calibration for regression or top-label calibration for classification that require using either the predicted CDF or indicator functions of $\hat{y}$ and y.

independent of θ and ϕ. If q_ϕ is minimized perfectly *including* the learnable prior distribution, then the bound becomes equal to the JS divergence. Note that this has a similar form to the adversarial approach except that it is a min problem and thus forms a min-min problem. A flow-based variant [Cho et al., 2022] provides an upper bound that only depends on optimizing the prior.

2.2.3 Non-Parametric Geometric Divergences

Geometric divergences (e.g., Wasserstein, Sinkhorn or MMD) vary with invertible transformations of the space. Intuitively, they depend on the distances in the space rather than ratios of densities as in information-theoretic divergences. One natural approach is to compute the distance between the domain distribution means. However, the means having a distance of zero is only necessary but not sufficient condition for the distributions to be equal. The maximum mean discrepancy (MMD) finds a function of random variables that maximizes the expectation between the domain distributions. While the function class could be a set of neural networks as in MMD-GAN Li et al. [2017], the most commonly used class of functions is a reproducing kernel hilbert space (RKHS) Gretton et al. [2012]. The MMD can be solved exactly when comparing empirical distributions, i.e., batches of samples from each domain. Thus, this empirical MMD can be used as a plug-in estimator of the distribution-level MMD:

$$\mathrm{D}^2_{\mathrm{MMD}}(\theta) \approx \widehat{\mathrm{D}}^2_{\mathrm{MMD}}(\theta)^{\cdot} = \|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\mathcal{H}} \tag{6}$$

$$= \widehat{\mathbb{E}}[\mathcal{K}(\mathbf{z}_1, \mathbf{z}_1)] - 2\widehat{\mathbb{E}}[\mathcal{K}(\mathbf{z}_1, \mathbf{z}_2)] + \widehat{\mathbb{E}}[\mathcal{K}(\mathbf{z}_2, \mathbf{z}_2)]\,, \tag{7}$$

where $\mathcal{H}$ is an RKHS with kernel $\mathcal{K}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ are the empirical (sample-based) means of domain 1 and 2 respectively in $\mathcal{H}$, and the expectations are based on unbiased sample averages [Gretton et al., 2012]. This can be seen as a generalization of comparing the empirical mean of the two distributions but using the implicit infinite dimensional space of a RKHS. One of the challenges is that this scales quadratically in the number of samples in the batch and thus cannot be computed for very large batches. Additionally, the performance can be sensitive to the kernel bandwidth parameter, which can be non-trivial to select in practice.

Another geometric divergence is based on Wasserstein distance. The Wasserstein-1 distance Panaretos and Zemel [2019] is defined optimal transport cost between the domain distributions using the the cost function $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$. The Wasserstein-1 between two empirical distributions (i.e., samples) can be computed by solving a linear program. Recently, linear program neural network layers have been proposed, which could be used to approximate it Mazouz et al. [2022]. However, solving a linear program for every batch of training samples is likely too expensive. In practice, an approximation to the Wasserstein distance based on an entropy-regularized optimal transport problem is often used. For this approximation, the Sinkhorn algorithm Cuturi [2013], which only requires matrix-vector multiplications, is often used since it has a complexity of $O(m^2 N_{\text{iter}})$ where m is the dimensionality and N_{iter} is the max number of Sinkhorn iterations. This approximation can be written as a regularized optimization problem Cuturi [2013]:

$$\begin{aligned} \hat{D}_{\text{SINK}}(\theta) := \Big(\mathbb{E}_{\hat{\pi}_\lambda}\left[c(\mathbf{z}_1, \mathbf{z}_2)\right] \\ \text{s.t. } \hat{\pi}_\lambda := \arg\min_{\hat{\pi}\in\Pi} \mathbb{E}_{\hat{\pi}}\left[\|\mathbf{z}_1 - \mathbf{z}_2\|_2\right] \\ + \lambda \mathsf{H}(\hat{\pi})\Big)^{\lambda\to 0} \approx D_{\mathrm{W}_1} \end{aligned} \tag{8}$$

where $\hat{\pi}_\lambda := \arg\min_{\hat{\pi}\in\Pi} \mathbb{E}_{\hat{\pi}}[\|\mathbf{z}_1 - \mathbf{z}_2\|_2] + \lambda\mathrm{H}(\hat{\pi})$ and where $\hat{\pi}(\mathbf{z}_1, \mathbf{z}_2) \in \Pi$ is the empirical coupling distribution between samples from each domain, Π corresponds to the set of joint discrete probability distributions over $\mathbf{z}_1$ and $\mathbf{z}_2$ whose marginals are $p_\theta(\mathbf{z}|\mathrm{d}{=}1)$ and $p_\theta(\mathbf{z}|\mathrm{d}{=}2)$, respectively, and $\overset{\lambda\to 0}{\approx}$ means that it approaches the true Wasserstein-1 distance as λ goes to zero. Note that this has two approximations. First, it compares a batch samples from each domain rather than the population-level distributions. Second, if $\lambda > 0$, then it forms an approximation to the Wasserstein-1 distance. While the Sinkhorn algorithm improves the computational complexity significantly, the algorithm is still at least quadratic in the number of samples in the batch and thus, like MMD, is difficult to apply for a large number of samples.

Previously mentioned geometric divergences have some problem, first MMD suffers from flat geometry, which eventually result into vanishing gradient Feydy et al. [2019b]. Also, vanilla OT causes dimension collapse on source mapping due to $\hat{D}_{SINK}(z_1, z_1) \neq 0$ by entropic regularization, thus it will introduce bias solution. Therefore, Sinkhorn divergence addresses above problem by interpolating between MMD and Sinkhorn with additional auto correlation term to prevent bias.

$$\begin{aligned} \hat{D}_{SINKD}(\theta) \overset{\text{def.}}{=} \hat{D}_{SINK}(z_1, z_2) - \frac{1}{2}\hat{D}_{SINK}(z_1, z_1) \\ - \frac{1}{2}\hat{D}_{SINK}(z_2, z_2) \end{aligned} \tag{9}$$

$$\hat{D}_{W_1} \xleftarrow{\varepsilon\to 0} \hat{D}_{SINKD}(\theta) \xrightarrow{\varepsilon\to+\infty} \hat{D}^2_{MMD}(\theta) \tag{10}$$

3 NORMALIZED GEOMETRIC DIVERGENCES FOR EVALUATING DM METHODS

One of the key challenges with comparing DM methods is properly evaluating how well the DM constraint was satisfied. Ideally, we would measure the divergence of the latent domain distributions. However, even estimating distribution divergences is known to be a challenging problem in its own right. While the adversarial and VAE-based methods could provide bounds on information theoretic divergences, they would require training an auxiliary model at test time to evaluate each method. Thus, we focus on the non- parametric divergences MMD and Sinkhorn that can be estimated with only samples. However, there is one key challenge with these geometric divergences when comparing across diverse methods. The scale of the latent distribution can significantly affect the absolute MMD or Sinkhorn divergence estimate because geometric divergences are highly sensitive to scale. This is a problem if the latent space is learned since the latent space scale is arbitrary. Thus, comparing methods using MMD or Sinkhorn directly would be unfair. To overcome this, we propose a simple approach based on applying ZCA whitening of the latent space before measuring the divergence. This ensures that the scale of the latent distributions is removed

4 RELATED WORK

Calibration Even though many deep learning models achieve high predictive performance, they often produce unreliable predictions due to a lack of calibration. Most deep learning models tend to be overconfident, as indicated by spiking posterior distributions Guo et al. [2017a]. Several factors contribute to this issue, including over-parameterized networks, insufficient regularization, limited data, and imbalanced label distributions Guo et al. [2017a]. There has been extensive research on calibration in both classification Bröcker [2009], Kull et al. [2017], Naeini et al. [2015b], Platt et al. [1999b], Dwork et al. [2021], Hébert-Johnson et al. [2018], Pleiss et al. [2017] and regression tasks Ziegel and Gneiting [2014], Kuleshov et al. [2018], Gneiting and Ranjan [2013], Song et al. [2019a], Zhao et al. [2020]. However, much of the community's focus has been on binary classification settings Karandikar et al. [2021], Vaicenavicius et al. [2019a], Bohdal et al. [2021], Platt et al. [1999a], Guo et al. [2017a]. Recently, Marx et al. [2024a] extended calibration into the distribution matching framework by leveraging the Maximum Mean Discrepancy (MMD)-based metric. This work unified recent advances in calibration across classification and regression tasks Kuleshov et al. [2018], Sahoo et al. [2021], Gneiting and Ranjan [2013], Zhao et al. [2021], Pessach and Shmueli [2022], Song et al. [2019a], Zhao et al. [2020], Luo et al. [2022]. Among the various calibration methods, our work focuses on individual calibration Zhao et al. [2020] conditioned on the variable $\mathbf{x}$.

Fairness Fairness in machine learning has garnered significant attention from the research community, with the primary goal of ensuring that machine learning models do not exhibit bias toward specific groups or individuals. Fairness algorithms are broadly categorized into two types: group fairness and individual fairness. Group fairness emphasizes equitable treatment across predefined demographic groups (e.g., male and female), while individual fairness ensures that similar individuals are treated similarly. To mitigate bias in machine learning models, researchers have proposed three primary strategies: preprocessing Creager et al. [2019], Lu et al. [2020], in-processing Chen and Wu [2020], Chiu et al. [2024], and post-processing Dwork et al. [2012], Hardt et al. [2016]. Preprocessing techniques modify the data before training, such as through normalization, relabeling, or reweighting. Post-processing methods adjust model outputs after training, typically at test time. In contrast, in-processing approaches impose fairness constraints during the training phase and have gained significant attention due to their ability to directly influence model behavior.

Our work focuses on in-processing methods, which are particularly relevant for enforcing fairness constraints during training. Prior studies in this area have primarily concentrated on specific applications or methods, often restricting their analysis to either latent space or logit space techniques. For instance, recent benchmark efforts have predominantly explored latent space approaches without extending their analysis to logit space methods Han et al. [2023b]. Additionally, these works often fail to provide a comprehensive comparison across different fairness methods or applications. In contrast, our study systematically evaluates in-processing methods by leveraging fairness techniques in both latent and logit spaces. We incorporate distribution-matching constraints and then evaluate their effectiveness using both information-theoretic and geometric divergence metrics. Consequently, we have a more holistic understanding of the trade-offs between different fairness methods. By addressing these gaps, our work provides a more comprehensive benchmark for group fairness methods compared to existing literature.

Domain Adaptation Domain adaptation seeks to enhance model generalization on out-of-distribution data. In this work, we focus on closed-set unsupervised domain adaptation, where the source and target domains share the same label space, but only the source domain is labeled.

Early methods aligned source and target feature distributions using statistical losses—for example, integrating a multi-kernel Maximum Mean Discrepancy (MMD) loss into deep neural networks Long et al. [2015]. Subsequent works refined these techniques [Long et al., 2017, Bousmalis et al., 2016] or introduced related MMD variants [**?**Zellinger et al., 2017, Kang et al., 2019]. In parallel, adversarial approaches have gained traction due to its flexibility and effectiveness. By incorporating a domain discriminator that distinguishes between source and target features, feature extractors can be trained to deceive the discriminator, thereby promoting domain-invariant representations [Ajakan et al., 2014, Ganin and Lempitsky, 2015, Ganin et al., 2016b, Tzeng et al., 2017]. Although less common, recent studies have also leveraged Sinkhorn divergences for domain adaptation [Pandya et al., 2025, Han et al., 2025], offering a promising alternative that efficiently aligns latent spaces via regularized optimal transport.

Many previous domain adaptation benchmarks evaluate models with dedicated designs that are intrinsically tied to specific divergence measures and task formulations [Lalou et al., 2025, **?**]. In contrast, our work introduces a unified distribution matching framework that employs a generalized network architecture across all experiments. By keeping the architecture fixed, we interchange different divergence measures (e.g., Sinkhorn, adversarial, MMD, and variational methods) and systematically assess their relationship with domain adaptation performance under uniform experimental conditions.

5 EMPIRICAL COMPARISON OF DM METHODS ACROSS TRUSTWORTHY ML TASKS

In this section, we focus on answering two major questions for each task.

RQ 1: What are the relationship between intrinsic metric (MMD, Sinkhorn) and task specific metric (DP (Fairness), ECE (Calibration), target accuracy (Domain Adaptation)?

RQ 2: Which DM method should we use for each TML task?

5.1 EXPERIMENT SETUP

We tuned the hyper parameters using a TPES sampler Bergstra et al. [2011] to find the best model, and used early stopping with tracking validation loss. A detailed experiment setup can be found in the appendix. For the calibration and fairness tasks, we used the ADULT dataset Becker and Kohavi [1996], considering gender (male and female) as the sensitive attribute and classifying income > 55k. For the domain adaptation task, we used the MNIST $\rightarrow$ USPS dataset Deng [2012], Hull [1994]. When evaluating Sinkhorn divergence and MMD, we applied ZCA whitening to the latent space when using latent space-based methods. However, we did not apply ZCA whitening Kessy et al. [2018] to the logit space since it is already a constrained space. We used the default epsilon (entropic regularization parameter) for Sinkhorn divergence from the GeomLoss library Feydy et al. [2019b]. For MMD, ongoing research seeks to determine the optimal bandwidth, as MMD is highly sensitive to this parameter. Initially, we applied the most common approach, median heuristicGarreau et al. [2017], but it did not perform well. Therefore, we experimented with bandwidth values of [1, 5, 10, 15, 20, 25] and selected the bandwidth that resulted in the highest MMD.

5.2 CALIBRATION

We follow an individual calibration approach, as described in Marx et al. [2024a]. While prior work primarily used the Maximum Mean Discrepancy (MMD) method, we extended the study by incorporating both the Sinkhorn divergence and an adversarial method. Since no prior work has applied adversarial techniques in this context, we implemented a GAN-based method designed to match the predicted distribution to the target ground-truth distribution.

For calibration, we applied temperature scaling as a post-hoc calibration technique Hinton [2015] and used the Expected Calibration Error (ECE) as our primary evaluation metric Naeini et al. [2015a].

Definition 2. *Expected Calibration Error (ECE) measures the discrepancy between model confidence and accuracy.*

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathbb{E}\left[\mathbb{I}\left(\hat{y} = y\right) \mid \hat{\mathbf{q}} \in B_m\right] - \mathbb{E}\left[\hat{\mathbf{q}} \mid \hat{\mathbf{q}} \in B_m\right] \right|, \tag{11}$$

where B_m denotes the set of samples in the m-th confidence bin, $|B_m|$ is the number of samples in bin m, n is the total number of samples, $\hat{\mathbf{q}}$ is predicted probability (confidence), and $\hat{y}$ is predicted label

Research Question 1 (RQ1): What is the relationship between intrinsic metrics and task-specific metrics?

Observation 1: The Sinkhorn Divergence exhibits a negative correlation with both ACC and ECE while MMD exhibits no strong correlation.

To understand the impact of distribution matching (DM) on calibration, we must examine the definition of ECE. The goal of DM in calibration is to align the predicted distribution q with the true distribution $p(y|x)$. However, perfect distribution matching tends to produce overconfident predictions, resulting in higher accuracy (ACC) but also increased ECE, as shown in Figure 2. In Figure 2, we observe lower Sinkhorn (strict DM) value have higher ACC and ECE with a negative correlation between the Sinkhorn divergence and accuracy, alongside a negative correlation with ECE.

Interestingly, this trend is less evident with MMD. This discrepancy can be attributed to the higher entropic regularization factor in MMD, which makes it a noisier estimator compared to the Sinkhorn divergence.

10.

Observation 2: There is a trade-off between ACC and ECE.

Extensive research in fairness has investigated trade off between Demographic Parity (DP) with ACC Han et al. [2023a], Plecko and Bareinboim [2024], Gong et al. [2024]. However, trade off between ACC and ECE haven't been

Method	ACC	Sink	MMD	ECE
Calibration (MMD)	**0.853 ± 0.0004586**	0.324 ± 0.00008116	**0.001776 ± 0.00001326**	**0.07271 ± 0.0004055**
Calibration (Sink)	**0.853 ± 0.0004072**	0.3231 ± 0.0001792	0.001815 ± 0.0001802	0.08282 ± 0.0008516
Calibration (Adv)	0.8509 ± 0.0008675	**0.3224 ± 0.0001652**	0.00183 ± 0.0001401	0.08994 ± 0.0009532

Table 1: Performance comparison of different calibration methods.

Method	ACC	Sink	MMD	ECE
Post Hoc Calibration (MMD)	0.853 ± 0.0004586	**0.3473 ± 0.000075**	**0.002384 ± 0.00001373**	0.08139 ± 0.0003872
Post Hoc Calibration (Sink)	**0.853 ± 0.0004072**	0.3492 ± 0.0001482	0.002494 ± 0.00001943	0.04953 ± 0.003339
Post Hoc Calibration (Adv)	0.8509 ± 0.0008675	0.3501 ± 0.0002244	0.002541 ± 0.00001443	**0.04865 ± 0.001214**

Table 2: Performance comparison of post hoc calibration methods.

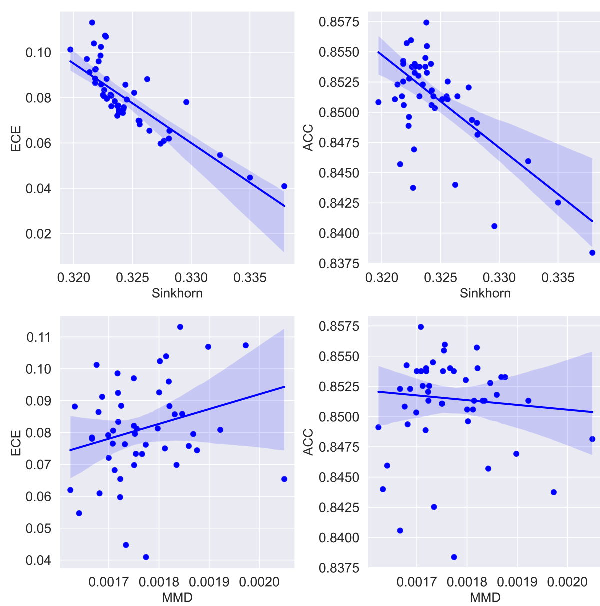

Figure 2: Relation between Sinkhorn Divergence and MMD with Accuracy and ECE on Calibration

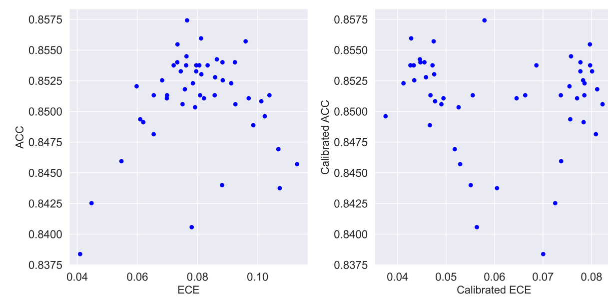

Figure 3: ACC vs ECE Trade off (Left), Calibrated ACC vs ECE Trade off (Right) for Calibration

thoroughly explored in calibration domain. Recent work revealed that there is negative correlation between ECE and ACC on high ACC model Tao et al. [2023a]. We therefore, further investigated on this topic. As we can observe on figure 3, initially, as ACC increases, ECE also increases. However, after reaching a certain point, ACC starts to decrease as ECE continues to increase. The reason behind positive correlation between ACC and ECE is when ACC is low, predictive confidence is also low, resulting in a small gap between confidence and accuracy, and as ACC increase, confidence increase, thus ECE increase as well Si et al. [2022]. However, beyond a certain confidence threshold, accuracy begins to decline due to overfitting confidence on incorrect predictions. Therefore, we can observe that **ACC exhibits a positive correlation with ECE in low-accuracy models and a negative correlation with ECE in high-accuracy models.** Our findings align with the observations in Tao et al. [2023a], which reported a negative correlation between ACC and ECE in high-accuracy models. Additionally, our results provide new insight by highlighting that low-accuracy models tend to display a positive correlation between ACC and ECE.

Observation 3: Post hoc calibration removes the ACC and ECE trade off. Ideally, Temperature Scaling should reduce ECE without effecting ACC, thus remove the ACC and ECE trade off Guo et al. [2017b]. Our result also follow this conjecture. As shown on figure 3, we can observe that equal ACC model exhibits different ECE, thus removed trade off.

RQ 2: Which DM method should we use for Calibration task?

We hypothesize that due to the strong correlation between Sinkhorn and both ACC and ECE as shown on figure 2, Sinkhorn tends to overfit, ultimately leading to an increase in ECE compared to MMD before post hoc calibration. However, applying post-hoc calibration significantly reduces ECE for both Sinkhorn and adversarial methods, while ECE increases for MMD. This phenomenon can be explained by a strongly regularized calibration method during training compresses logit distributions and removes sample difficulty information, thereby limiting the potential improvement achievable through post-hoc calibration Wang et al. [2021]. Therefore, several studies recommend using both training-time and post-hoc calibration as a unified framework rather than relying solely on individual methods. **In conclusion, we recommend practitioner to use Sinkhorn method with post hoc calibration.**

6 FAIRNESS

Most fairness benchmark papers Han et al. [2023a], **?**], Reddy [2022a] focus on fair representation learning, which we refer to as latent-based methods. However, there is a lack of prior work studying logit-based approaches Chung et al. [2024b]. In this paper, we compare distribution matching using both logit-based and latent-based methods using Sinkhorn, MMD, adversarial, and VAUB. However, VAUB is latent based method, so we did not compare VAUB with logit base method Gong et al. [2024]. Interestingly, definition of fairness metric "demographic parity" is closely related to distribution matching that strong DM will result into lower DP. Therefore, in this section, we are going to show how DM help fairness task.

Definition 3. ***Demographic Parity*** *measures discrepancy of true positive rate between different domain.*

$$|p(\hat{y} = 1|d = 1) - p_\theta(\hat{y} = 1|d = 2)| \tag{12}$$

Throughout the experiment, we use ratio instead of absolute difference following Torchmetric implementation

$$\frac{\min_d\ p\left(\hat{y} = 1 \mid d\right)}{\max_d\ p\left(\hat{y} = 1 \mid d\right)}. \tag{13}$$

RQ 1: What are the relationship between intrinsic metric and task specific metric?

Observation: For logit based method, both Sinkhorn divergence and MMD exhibits a negative correlation with demographic parity (DP) with positive correlation with ACC while no strong correlation observe on latent based method.

For the latent space method, as shown in Figure 4, we observe that Sinkhorn and MMD do not exhibit a strong correlation with ACC and DP, thus failing to provide a meaningful trend for latent based method. In contrast, as shown on figure 5, logit-based methods demonstrate negative correlation between Sinkhorn and MMD with DP, which indicates that DM is effective on fairness task. For ACC, we observe positive correlation between MMD and Sinkhorn. Therefore, there is inherent trade off between ACC and DP.

RQ 2: Which DM method should we use for Fairness task

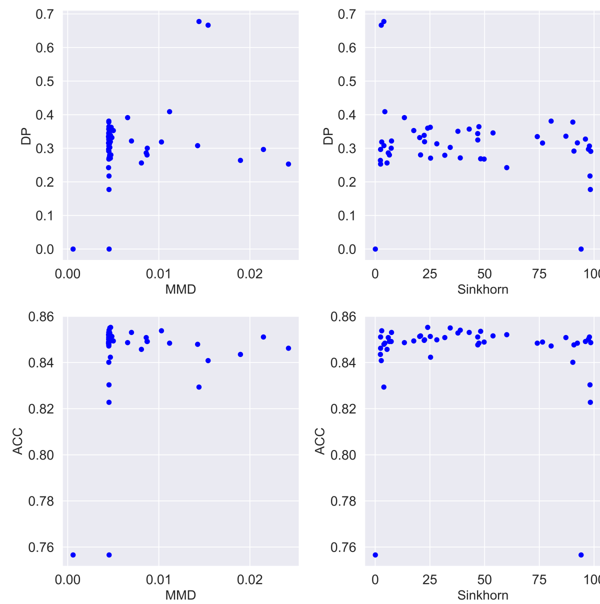

Figure 4: Relation etween Sinkhorn Divergence and MMD with Accuracy and DP for latent based fairness

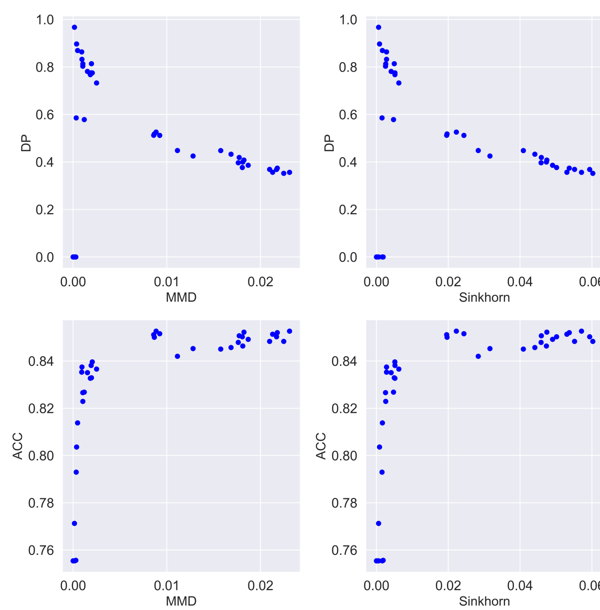

Figure 5: Relation etween Sinkhorn Divergence and MMD with Accuracy and DP for logit based fairness

Notably, most logit-based methods outperform latent-based methods, yielding higher accuracy and DP, as shown in Tables 3 and 4. This can be explained by the fact that logit-based methods exhibit a stronger correlation between the intrinsic metric and the task-specific metric compared to latent-based methods, as shown in Figures 4 and 5. Therefore, using logit-based methods is more effective. Interestingly, Sinkhorn achieves almost perfect DP with a trade-off

Method	ACC	DP	Sink	MMD
Fairness Latent based (MMD)	0.8468 ± 0.001215	0.317 ± 0.009764	71.834 ± 3.762	**0.004489** ± 0.0000184
Fairness Latent based (Sink)	0.8447 ± 0.0006408	**0.4493** ± 0.005768	**0.4479** ± 0.1018	0.01706 ± 0.001062
Fairness Latent based (Adv)	0.847 ± 0.0007036	0.3085 ± 0.01354	50.456 ± 4.761	0.004562 ± 0.0000106
Fairness Latent based (VAUB)	**0.8578** ± 0.0008785	0.2968 ± 0.01286	56.066 ± 1.631	0.00475 ± 0.00002788

Table 3: Fairness latent based methods performance comparison.

Method	ACC	DP	Sink	MMD
Fairness Logit based (MMD)	0.8489 ± 0.001146	0.3457 ± 0.00962	0.06427 ± 0.001746	0.02389 ± 0.0006595
Fairness Logit based (Sink)	0.8187 ± 0.001124	**0.9181** ± 0.01749	**0.001025** ± 0.0000754	**0.0003324** ± 0.00002196
Fairness Logit based (Adv)	**0.852** ± 0.0004109	0.306 ± 0.005739	0.06563 ± 0.00173	0.02712 ± 0.0007715

Table 4: Fairness Latent based methods performance comparison.

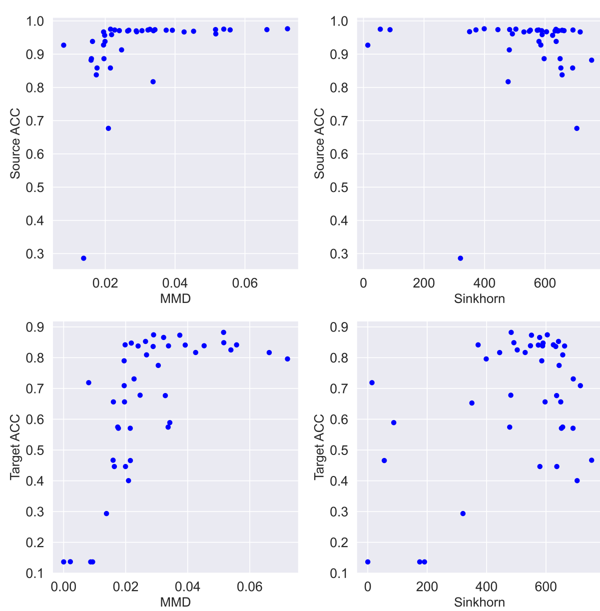

Figure 6: Relation between Sinkhorn Divergence and MMD with Source Accuracy and Target Accuracy on Domain Adaptation MNIST →USPS

in accuracy. This is because the Sinkhorn method tends to overfit to distribution matching, resulting in lower accuracy. **Therefore, we recommend practitioners use logit-based methods rather than latent-based methods. Specifically, if fairness is a priority, we suggest using Sinkhorn. For a well-balanced trade-off between accuracy and DP, MMD-based method is recommended.**

7 DOMAIN ADAPTATION

In this paper, we focus on unsupervised domain adaptation setting where we do not have access to the target label. Wilson and Cook [2020]. We use Sinkhorn based method Courty et al. [2014], MMD based method Tzeng et al. [2014], and adversarial based method. Ganin et al. [2016a].

RQ 1: What is the relationship between the intrinsic metric and the task-specific metric?

Observation: Strict DM is not always beneficial for domain adaptation.

In domain adaptation, Ben-David et al. [2006b] provides a useful bound on target error in terms of source error using the $\mathcal{A}$ distance, suggesting that a good representation should have both low source error and low $\mathcal{A}$ distance between source and target distributions. However, computing the $\mathcal{A}$ distance in practice is often impractical, so we instead leverage geometric divergence to measure the distance between two distributions. Interestingly, our experimental results contradict the direct implication of the above theoretical bound. As shown in figure 6, a low geometric divergence does not necessarily lead to high target accuracy. Conversely, as geometric divergence increases up to a certain point, the target accuracy also increases; beyond that point, the target accuracy begins to decrease. **These results provide a new perspective on domain adaptation: enforcing strict distribution matching may not always be optimal, and allowing a certain gap between the source and target distributions can be more beneficial.** In contrast to the trend observed in target ACC, the source ACC is not highly sensitive to geometric divergence, yielding similar source ACC values across the different geometric divergence settings.

RQ 2: Which DM method should we use for domain adaptation tasks?

As shown in figure 6, Sinkhorn exhibits a stronger correlation with both target ACC and source ACC, indicating that Sinkhorn is an effective DM metric. Moreover, as presented in table 5, Sinkhorn achieves both the highest source ACC

Method	ACC (Source)	ACC (Target)	Sink	MMD
Domain Adaptation (MMD)	0.9439 ± 0.005122	0.6784 ± 0.02159	$\mathbf{312.998} \pm 16.392$	0.07625 ± 0.001505
Domain Adaptation (Sink)	$\mathbf{0.9736} \pm 0.0009812$	$\mathbf{0.8494} \pm 0.004516$	617.789 ± 7.445	$\mathbf{0.0299} \pm 0.0015$
Domain Adaptation (Adv)	0.9714 ± 0.001038	0.6511 ± 0.02468	419.11 ± 30.62	0.1147 ± 0.008344
Domain Adaptation (VAUB)	0.9681 ± 0.001431	0.5793 ± 0.01856	512.636 ± 4.163	0.06615 ± 0.002131

Table 5: Comparison of Domain Adaptation methods based on ACC (Source and Target), Sink, and MMD metrics.

and the highest target ACC, with a significant margin of roughly 20% over the other methods. **Therefore, we recommend that practitioners leverage the Sinkhorn method for domain adaptation tasks.**

8 CONCLUSION

This study explores the application of distribution matching (DM) to three TML tasks: calibration, fairness, and domain adaptation. Through extensive experiments, we show how intrinsic metrics relate to task-specific metrics in each TML task. Our findings reveal a strong correlation between Sinkhorn and various task-specific metrics, indicating that Sinkhorn is an effective DM regularizer. This trend also leads to better performance across these TML tasks, making Sinkhorn an appealing "go-to" approach for DM. Additionally, our results provide meaningful insights, which build interesting connection between previous work. For instance, recent work on calibration revealed that strong ACC model exhibits negative correlation between ACC and ECE, and our results further provide that on weak ACC model, it exhibits positive correlation. Additionally, prior work on fairness typically focuses on latent-based (representation learning) methods, but our experiments demonstrate that logit-based methods can outperform latent-based methods. Likewise, previous domain adaptation research relies on the theoretical bound involving the $\mathcal{A}$ distance, suggesting that a closer match between distributions should yield higher target accuracy. However, our empirical findings show that strict distribution matching can harm performance for domain adaptation. Instead, allowing some gap between the source and target distributions can improve performance. We hope these insights will facilitate the development of more effective DM methods for a variety of TML tasks.

Author Contributions

Briefly list author contributions. This is a nice way of making clear who did what and to give proper credit. This section is optional.

H. Q. Bovik conceived the idea and wrote the paper. Coauthor One created the code. Coauthor Two created the figures.

Acknowledgements

Briefly acknowledge people and organizations here.

All acknowledgements go in this section.

References

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Shun-ichi Amari. Divergence, optimization and geometry. In *International conference on neural information processing*, pages 185–193. Springer, 2009.

Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

Barry Becker and Ronny Kohavi. Adult. *UCI Machine Learning Repository*, 10:C5XW20, 1996.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006a. URL `https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf`.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006b.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*, 2021.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/45fbc6d3e05ebd93369ce542e8f2322d-Paper.pdf`.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

Mingliang Chen and Min Wu. Towards threshold invariant fair classification. In *Conference on Uncertainty in Artificial Intelligence*, pages 560–569. PMLR, 2020.

Ching-Hao Chiu, Yu-Jen Chen, Yawen Wu, Yiyu Shi, and Tsung-Yi Ho. Achieve fairness without demographics for dermatological disease diagnosis. *Medical Image Analysis*, 95:103188, 2024.

Wonwoong Cho, Ziyu Gong, and David I. Inouye. Cooperative distribution alignment via jsd upper bound. In *Neural Information Processing Systems (NeurIPS)*, dec 2022.

Hao-Wei Chung, Ching-Hao Chiu, Yu-Jen Chen, Yiyu Shi, and Tsung-Yi Ho. Toward fairness via maximum mean discrepancy regularization on logits space. *arXiv preprint arXiv:2402.13061*, 2024a.

Hao-Wei Chung, Ching-Hao Chiu, Yu-Jen Chen, Yiyu Shi, and Tsung-Yi Ho. Toward fairness via maximum mean discrepancy regularization on logits space. *arXiv preprint arXiv:2402.13061*, 2024b.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/creager19a.html`.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019a.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019b.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016a.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016b.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. 2013.

Ziyu Gong, Ben Usman, Han Zhao, and David I. Inouye. Towards practical non-adversarial distribution matching. In **International Conference on Artificial Intelligence and Statistics (AISTATS)**, May 2024.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017a.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017b.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. *arXiv preprint arXiv:2306.09468*, 2023a.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. *arXiv preprint arXiv:2306.09468*, 2023b.

Yan Han, Ailin Hu, Qingqing Huang, Yan Zhang, Zhichao Lin, and Jinghua Ma. Sinkhorn divergence-based contrast domain adaptation for remaining useful life prediction of rolling bearings under multiple operating conditions. *Reliability Engineering & System Safety*, 253:110557, 2025.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pages 623–631. PMLR, 2017.

Yanis Lalou, Théo Gnassounou, Antoine Collas, Antoine de Mathelin, Oleksii Kachaiev, Ambroise Odonnat, Alexandre Gramfort, Thomas Moreau, and Rémi Flamary. Skada-bench: Benchmarking unsupervised domain adaptation methods with realistic validation on diverse modalities, 2025. URL `https://arxiv.org/abs/2407.11676`.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/long15.html`.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks, 2017. URL `https://arxiv.org/abs/1605.06636`.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202, 2020.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: metrics and recalibration. In *Uncertainty in Artificial Intelligence*, pages 1286–1295. PMLR, 2022.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 2023.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36, 2024a.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36, 2024b.

Rayan Mazouz, Karan Muvvala, Akash Ratheesh Babu, Luca Laurenti, and Morteza Lahijanian. Safety guarantees for neural network dynamic systems via stochastic barrier functions. *Advances in Neural Information Processing Systems*, 35:9672–9686, 2022.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015a.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015b.

Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.

Sneh Pandya, Purvik Patel, Brian D. Nord, Mike Walmsley, and Aleksandra Ćiprijanović. Sidda: Sinkhorn dynamic domain adaptation for image classification with equivariant neural networks, 2025. URL `https://arxiv.org/abs/2501.14048`.

Mdel C Pardo and Igor Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, 49(7):1860–1867, 2003.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55 (3):1–44, 2022.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999a.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999b.

Drago Plecko and Elias Bareinboim. Fairness-accuracy trade-offs: A causal perspective. *arXiv preprint arXiv:2405.15443*, 2024.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Yu Qiao and Nobuaki Minematsu. f-divergence is a generalized invariant measure between distributions. In *INTERSPEECH*, pages 1349–1352, 2008.

Charan Reddy. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. 2022a.

Charan Reddy. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. 2022b.

Roshni Sahoo, Shengjia Zhao, Alyssa Chen, and Stefano Ermon. Reliable decisions with threshold calibration. *Advances in Neural Information Processing Systems*, 34: 1831–1844, 2021.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. Re-examining calibration: The case of question answering. *arXiv preprint arXiv:2205.12507*, 2022.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019a.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019b.

Linwei Tao, Younan Zhu, Haolan Guo, Minjing Dong, and Chang Xu. A benchmark study on calibration. *arXiv preprint arXiv:2308.11838*, 2023a.

Linwei Tao, Younan Zhu, Haolan Guo, Minjing Dong, and Chang Xu. A benchmark study on calibration. *arXiv preprint arXiv:2308.11838*, 2023b.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019a.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd international conference on artificial intelligence and statistics*, pages 3459–3467. PMLR, 2019b.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820, 2021.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR, 2019.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. *Journal of machine learning research*, 23(340):1–49, 2022.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR, 2020.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324, 2021.

Johanna F Ziegel and Tilmann Gneiting. Copula calibration. 2014.