FAIR REPRESENTATION LEARNING BEYOND EUCLIDEAN GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair representation learning enforces that different demographic groups follow similar distributions in representation space, with similarity measured by a distributional discrepancy computed in Euclidean coordinates. This implicitly assumes the representation space is flat, which is an assumption violated whenever representations are embedded in curved spaces such as hyperbolic space, where the ambient Euclidean distance can differ arbitrarily from the geodesic distance the model actually uses. We show the mismatch changes the conclusion rather than the constant: two group distributions can be nearly matched under a Euclidean discrepancy while remaining well separated on the underlying manifold, so the resulting fairness guarantees do not transfer. We therefore measure discrepancy intrinsically, on Hadamard manifolds, non-positively curved spaces that include Euclidean and hyperbolic geometry as special cases and whose unique geodesics make projection well defined. Our Cartan–Hadamard Sliced-Wasserstein (CHSW) distance aggregates one-dimensional Wasserstein distances along geodesic projections, avoiding full optimal transport on the manifold while respecting its curvature, and is differentiable and cheap enough to serve as a training regularizer. We prove that minimizing it bounds demographic parity violation, and show on standard benchmarks that it improves fairness substantially at competitive accuracy.

1 INTRODUCTION

Algorithmic decision-making systems increasingly govern critical aspects of modern life, including credit approval (Kozodoi et al., 2022), hiring decisions (Schumann et al., 2020), criminal sentencing (Dressel & Farid, 2018), and healthcare resource allocation (Obermeyer et al., 2019). While these AI-driven systems promise improved efficiency and scalability, growing empirical evidence reveals a troubling pattern: machine learning models often inherit and amplify societal biases embedded in historical data (Barocas et al., 2023), leading to systematic discrimination against marginalized groups (Mehrabi et al., 2021). These concerns have, in turn, spurred intensive research on algorithmic fairness (Mitchell et al., 2021; Siddique et al., 2023), aiming to mitigate bias in automated decision-making systems.

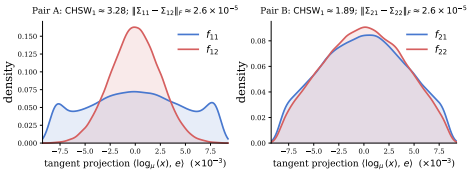


Figure 1: **Equal Euclidean parameter distances need not imply similar distributions on a manifold.** Two pairs of wrapped normals on the 3D Poincaré ball $\mathbb{B}^3$ ($K = -1$) share the mean $\mu = (0.9995, 0, 0)$ and have nearly identical Frobenius covariance gaps, $\approx 2.57 \times 10^{-5}$. However, their horospherical CHSW1 distances differ substantially: 3.28 versus 1.89. Each panel shows the KDE of the tangent projection $\langle \log \mu(x), e \rangle$ with $e = (1, 0, 0)$; CHSW$_1$ is computed independently by Monte Carlo horospherical slicing. Covariances in Appendix A.

Among the proposed approaches, fair representation learning (FRL) has emerged as a particularly compelling paradigm: it aims to learn an intermediate representation that reduces dependence on sensitive attributes while preserving predictive signal for downstream tasks, thereby enabling fairer decisions without sacrificing utility (Creager et al., 2019; Liu et al., 2022). The literature on fair representation learning is extensive, largely due to its versatility: once learned, such representations can be applied to multiple downstream tasks without requiring task-specific fairness interventions (Moyer et al., 2018; Sinigaglia et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, most existing FRL methods assume the learned representations lie in a Euclidean space, and assess the distributional similarity between groups (the operational notion of fairness) under Euclidean geometry. This choice is more consequential than it appears. By the manifold hypothesis, high-dimensional natural data, and the representations learned from it, tend to concentrate near a low-dimensional, curved manifold rather than filling a flat Euclidean space (Bengio et al., 2013; de Kruiff et al., 2025; Fefferman et al., 2016; Roweis & Saul, 2000). On such a manifold, Euclidean distance is the wrong ruler: it measures separation along straight lines through the ambient space, whereas a downstream predictor, or an adversary trying to recover the protected attribute, operates in the distance along the manifold. As shown in Figure 1 and Figure 4 (Appendix A), two group distributions can lie almost on top of each other under a Euclidean metric yet remain clearly separated on the underlying manifold. A fairness regularizer that drives the Euclidean gap to zero therefore only makes the groups look indistinguishable. Their group-specific structure survives along the manifold and stays recoverable, leaving the resulting fairness guarantees unreliable.

Building on the above observation, we develop a FRL framework that measures and enforces fairness directly on the manifold where the data live, rather than on a Euclidean stand-in for it. Doing so confronts two obstacles: an arbitrary manifold is analytically and computationally intractable, and even on a fixed manifold, comparing group distributions via optimal transport is expensive. Our framework is designed to dissolve both. We map inputs to latent representations on *Hadamard manifolds*, which retain the curvature that makes the geometry matter while offering enough structure to keep the quantities we need tractable (Bridson & Haefliger, 2013). On these spaces we introduce the Cartan–Hadamard Sliced-Wasserstein (CHSW) distance (Bonet et al., 2025), which compares group distributions through cheap one-dimensional geodesic slices, capturing curvature effects that Euclidean measures miss at comparable cost. Finally, we prove that minimizing a CHSW-based regularizer provably bounds demographic parity violation, so optimizing this geometry-aware quantity guarantees fairness under the standard, geometry-agnostic notion that practitioners report.

Contributions. Our main contributions are fourfold: **(1)** We propose the first FRL framework that operates beyond Euclidean geometry: we show that fairness assessed in a Euclidean latent space is geometrically blind, group distributions can look nearly identical under Euclidean metrics yet remain clearly separated on the manifold where the data live, and we correct these geometric biases that remain invisible to existing Euclidean methods. **(2)** We introduce a CHSW-based fairness regularization framework that captures true geometric distributional differences on manifolds while reducing them to tractable one-dimensional geodesic comparisons, recovering the geometry that Euclidean measures discard at essentially Euclidean cost and addressing a key challenge in manifold-based fairness assessment. **(3)** We prove that our method provably controls appropriate notions of demographic-parity violation with explicit bounds, providing the first fairness guarantees that respect the underlying geometric structure. **(4)** Experiments on standard fairness benchmarks show that our method maintains competitive accuracy while consistently improving fairness over SOTA baselines.

2 Related Work

Fair representation learning. Fair representation learning aims to minimize the correlation between learned latent representations and sensitive attributes while keeping the learned representations informative for downstream tasks. Existing methods can be broadly categorized into several directions. A large body of work adopts adversarial learning (Madras et al., 2018; Zhang et al., 2018). Another line of research focuses on variational autoencoders (Louizos et al., 2017). In addition, regularization-based methods enforce fairness by penalizing statistical dependence between learned representations and sensitive attributes using information-theoretic or distance-based measures (Dzakpasu et al., 2024; Oneto et al., 2020), and our work falls into this category. Recent works also explore geometry-aware and distributional methods by using optimal transport or representation alignment (Gaucher et al., 2023; Nguyen et al., 2025). However, these methods quantify distributional discrepancies within Euclidean space to assess fairness, while overlooking the manifold structure underlying the distributions. As a result, their fairness control can be inaccurate and less reliable.

Wasserstein-based fairness methods. Wasserstein distance has become a powerful tool in fairness literature for assessing distributional similarity between demographic groups (Gordaliza et al., 2019). Existing work applies Wasserstein distance to match distributions across demographic groups in the learned representation space (Wang et al., 2024; Xiong et al., 2024a;b). For example, some

methods enforce demographic parity by minimizing the transport cost between groups (Chzhen et al., 2020; Gaucher et al., 2023). Other methods introduce entropic regularization to improve computational efficiency (Jiang et al., 2020) or apply multi-marginal formulations to handle multiple groups (Chowdhary et al., 2024; Nguyen et al., 2025). These methods primarily study fairness in Euclidean representation or prediction spaces, where the chosen geometry may fail to capture distributional differences on an underlying manifold. Prior work has developed hyperbolic SW (Bonet et al., 2023a), SPD SW (Bonet et al., 2023b), and intrinsic SW on manifolds (Rustamov & Majumdar, 2023). Building on the CHSW construction of Bonet et al. (2025), we develop a computationally efficient fairness regularizer for manifold-valued representations.

3 PRELIMINARIES

This section introduces the problem formulation and the key concepts underlying our method.

3.1 PROBLEM FORMULATION

We consider a binary classification problem with a sensitive attribute. Let (X, S, Y) denote the random variables for input features, sensitive attribute, and binary label, respectively, with $X \in \mathcal{X}$, $S \in [A] := \{0, 1, \dots, A-1\}$ and $Y \in \{0, 1\}$. We are given a training dataset $\mathcal{D} = \{(x_i, s_i, y_i)\}_{i=1}^n$ drawn i.i.d. from a joint distribution $\mathbb{P}_{X,S,Y}$. Throughout, we focus on the binary case ($A = 2$), while our framework naturally extends to multi-class settings ($A > 2$).

Following the fair representation learning paradigm, we learn an encoder $f_\theta : \mathcal{X} \to \mathcal{M}$ that maps inputs to manifold-valued representations $Z = f_\theta(X) \in \mathcal{M}$. Here, $\mathcal{M}$ is chosen to be a Hadamard manifold. We also learn a predictor $h_\psi : \mathcal{M} \to [0, 1]$ that outputs $\hat{Y} = h_\psi(Z)$. The model is trained by minimizing:

$$\mathcal{L}(\theta, \psi) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{BCE}}(h_\psi(f_\theta(X_i)),\, Y_i) + \lambda\, \mathcal{R}_{\mathrm{fair}}(\{Z_i\}_{i=1}^n, \{S_i\}_{i=1}^n), \tag{1}$$

where $\ell_{\mathrm{BCE}}(\hat{y}, y) = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$ is the binary cross-entropy loss, $\lambda > 0$ is a balancing coefficient, and $\mathcal{R}_{\mathrm{fair}}$ is a fairness regularizer that measures distributional dissimilarity between demographic groups in the representation space.

3.2 HADAMARD MANIFOLDS

In this paper, we focus on Hadamard manifolds (a.k.a. Cartan-Hadamard manifolds), i.e., complete, simply connected Riemannian manifolds with non-positive sectional curvature (Bridson & Haefliger, 2013; Lang, 2012; Lee, 2006; Robbin & Salamon, 2022). This class includes pullback Euclidean spaces (with Euclidean space as a special case), hyperbolic spaces, and the manifold of symmetric positive-definite (SPD) matrices endowed with appropriate metrics (Bonet et al., 2025).

Definition 3.1 (Pullback Euclidean Space)**.** Let $(\mathcal{N}, \langle\cdot,\cdot\rangle)$ be a Euclidean space and $\phi : \mathcal{M} \to \mathcal{N}$ a diffeomorphism. The pullback metric on $\mathcal{M}$ is $g_x^\phi(u, v) = \langle \phi_{*,x}(u), \phi_{*,x}(v)\rangle$, under which $d_{\mathcal{M}}(x, y) = \|\phi(x) - \phi(y)\|$. Such pullback Euclidean manifolds are flat and form a subclass of Hadamard manifolds.

Definition 3.2 (Hyperbolic Space)**.** Hyperbolic space is a complete, simply connected Riemannian manifold with constant negative sectional curvature $K < 0$, and hence is a Hadamard manifold.

Definition 3.3 (Manifold of Symmetric Positive Definite Matrices)**.** Let $\mathbb{S}^d(\mathbb{R})$ be the space of real symmetric $d \times d$ matrices. The SPD manifold is $\mathbb{S}^d_{++}(\mathbb{R}) = \left\{ M \in \mathbb{S}^d(\mathbb{R}) : x^\top M x > 0,\ \forall x \in \mathbb{R}^d \setminus \{0\} \right\}$. It is a smooth manifold, and the tangent space at M can be identified with $\mathbb{S}^d(\mathbb{R})$.

3.3 CARTAN-HADAMARD SLICED-WASSERSTEIN

We use the CHSW distance because it lifts sliced optimal transport to Hadamard manifolds by averaging one-dimensional Wasserstein distances of projected measures, yielding a tractable yet geometry-aware discrepancy between distributions beyond the Euclidean setting.

Definition 3.4 (Cartan-Hadamard Sliced-Wasserstein)**.** Let $(\mathcal{M}, g)$ be a Hadamard manifold with o as its origin. Denote by λ_o the uniform distribution on $S_o = \{\, v \in T_o\mathcal{M} \;:\; \|v\|_o = 1 \,\}$. With $p \geq 1$, we define the *p-Cartan–Hadamard Sliced–Wasserstein distance* between $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$ as

$$\mathrm{CHSW}_p^p(\mu, \nu) = \int_{S_o} W_p^p\big(P^v_{\#}\mu,\ P^v_{\#}\nu\big)\, \mathrm{d}\lambda_o(v), \tag{2}$$

where $P^v : \mathcal{M} \to \mathbb{R}$ is a one-dimensional projection along direction v, defined as either (i) the *geodesic projection* onto $\gamma_v(t) = \exp_o(tv)$ (yielding GCHSW_p^p), or (ii) the *horospherical projection* via the Busemann function associated with γ_v (yielding HCHSW_p^p).

4 METHOD

This section introduces our manifold-aware FRL framework based on CHSW barycenter regularization, with fairness guarantees on both pullback Euclidean and general Hadamard manifolds.

4.1 CHSW BARYCENTER REGULARIZATION FOR FAIRNESS

In FRL, we learn an encoder $f_\theta : \mathcal{X} \to \mathcal{M}$ that maps inputs to a latent manifold $\mathcal{M}$, $Z = f_\theta(X) \in \mathcal{M}$. Distributional demographic parity at the representation level requires $Z \perp\!\!\!\perp S$, i.e., the representation is distributionally independent of demographic groups. Equivalently, the pushforward distributions of Z must be identical across groups. Let $P_{X|S=s}$ denote the conditional input distribution for group $s \in [A]$, and define the induced representation distributions on $\mathcal{M}$ by

$$\mu_s \triangleq (f_\theta)_{\#} P_{X|S=s}, \qquad s = 0, \ldots, A-1,$$

where $(f_\theta)_{\#}$ denotes the pushforward operator. Then $Z \perp\!\!\!\perp S$ is equivalent to $\mu_0 = \cdots = \mu_{A-1}$.

To promote demographic parity at the representation level, we regularize the encoder by aligning the group-wise pushforward distributions $\{\mu_s\}_{s=0}^{A-1}$ toward a common CHSW barycenter. Given weights $\pi_s > 0$ satisfying $\sum_{s=0}^{A-1} \pi_s = 1$, we define the CHSW barycenter regularizer by

$$U_{\mathrm{CHSW}}(\mu_0, \ldots, \mu_{A-1}) \triangleq \min_{\nu \in \mathcal{P}_p(\mathcal{M})} \sum_{s=0}^{A-1} \pi_s\, \mathrm{CHSW}_p^p(\mu_s, \nu),$$

where CHSW_p^p can be instantiated as either GCHSW_p^p or HCHSW_p^p depending on the manifold geometry and the chosen projection operator. The details of P^v used in the experiments are provided in Appendix E.3.4.

Let $\bar{\mu}$ be any barycenter solution, $\bar{\mu} \in \arg\min_{\nu \in \mathcal{P}_p(\mathcal{M})} \sum_{s=0}^{A-1} \pi_s\, \mathrm{CHSW}_p^p(\mu_s, \nu)$. Then the objective can be equivalently written as

$$U_{\mathrm{CHSW}}(\mu_0, \ldots, \mu_{A-1}) \;=\; \sum_{s=0}^{A-1} \pi_s\, \mathrm{CHSW}_p^p(\mu_s, \bar{\mu}). \tag{3}$$

Intuitively, minimizing U_{CHSW} pulls the group-conditional distributions $\{\mu_s\}_{s=0}^{A-1}$ toward a common target $\bar{\mu}$, reducing inter-group discrepancies and improving demographic parity.

4.2 PULLBACK EUCLIDEAN MANIFOLD

A widely used group fairness criterion for binary classification is the *demographic parity violation*. Under the binary sensitive attribute setting, it is defined as

$$\Delta_{\mathrm{DP}} = \Big|\mathbb{P}\big(\hat{Y} = 1 | S = 0\big) - \mathbb{P}\big(\hat{Y} = 1 | S = 1\big)\Big|.$$

While Δ_{DP} is intuitive, it depends on thresholding and the particular choice of downstream decision rule. To obtain a representation-level notion of demographic parity that is invariant to the decision head, we instead adopt its distributional formulation, where fairness reduces to matching the conditional representation distributions across demographic groups. Motivated by this perspective, we quantify the resulting deviation on manifolds using the CHSW distance and define:

$$\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta) \triangleq \max_{s,t \in [A]} \mathrm{CHSW}_p^p(\mu_s, \mu_t).$$

Intuitively, $\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta)$ captures the worst-case discrepancy between any two group-wise representation distributions. When CHSW_p is a metric on $\mathcal{P}_p(\mathcal{M})$ (e.g., on pullback Euclidean manifolds), it yields an exact characterization of representation-level distributional demographic parity.

As CHSW_p satisfies the triangle inequality on pullback Euclidean manifold (Bonet et al., 2025, Proposition 26), we derive the following upper bounds on $\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}$.

Lemma 4.1. *We have*

$$\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta) \le \frac{2^p}{\pi_{\min}} U_{\mathrm{CHSW}}, \qquad \pi_{\min} = \min_{s\in[A]} \pi_s > 0. \tag{4}$$

The proof of Lemma 4.1 is presented in Appendix C.1. Bonet et al. (Bonet et al., 2025, Proposition 28) established a comparison inequality between the Wasserstein distance W_p and CHSW_p on pullback Euclidean manifolds as shown in Proposition 4.2.

Proposition 4.2. *Let $(\mathcal{M}, g^\phi)$ be a pullback Euclidean manifold of dimension d. Fix $p \ge 1$ and $r > 0$, and let $B(o,r) \triangleq \{x \in \mathcal{M} : d_{\mathcal{M}}(x,o) \le r\}$ be the closed geodesic ball centered at $o \in \mathcal{M}$ with radius r. Then there exists a constant $C_{d,p,r} > 0$ such that for all $\mu, \nu \in \mathcal{P}_p(B(o,r))$,*

$$W_p^p(\mu,\nu) \le C_{d,p,r}\, \mathrm{CHSW}_p(\mu,\nu)^{\frac{1}{d+1}}. \tag{5}$$

Based on Proposition 4.2, we obtain Lemma 4.3.

Lemma 4.3. *Fix a score head $h : \mathcal{M} \to [0,1]$ and a threshold $\tau \in (0,1)$, and define $c_\tau(z) \triangleq \mathbf{1}\{h(z) \ge \tau\}$. Then the demographic parity violation of the classifier $c_\tau \circ f_\theta$ is $\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta) = \max_{s,t\in[A]} \left|\mathbb{P}\big(h(Z) \ge \tau \mid S = s\big) - \mathbb{P}\big(h(Z) \ge \tau \mid S = t\big)\right|$.*

***Assume** 1. **(Lipschitz continuous)** h is L-Lipschitz w.r.t. $d_{\mathcal{M}}$, i.e., $|h(z) - h(z')| \le L\, d_{\mathcal{M}}(z,z')\ \forall z, z' \in \mathcal{M}$;*

*2. **(Local anti-concentration at the threshold)** There exists a constant $\rho > 0$ such that for every group $s \in [A]$ and all $\varepsilon > 0$, $\mathbb{P}\big(|h(Z) - \tau| \le \varepsilon \mid S = s\big) \le 2\rho\varepsilon$;*

*3. **(Bounded support)** There exist $o \in \mathcal{M}$ and $r > 0$ such that $\mathrm{supp}(\mu_s) \subseteq B(o,r)$ for all $s \in [A]$.*

Then

$$\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta) \ \le\ 2\sqrt{2\rho\, L}\, C_{d,p,r}^{\frac{1}{2p}} \left(\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta)\right)^{\frac{1}{2(d+1)p^2}}, \tag{6}$$

where $C_{d,p,r} > 0$ is the constant in Proposition 4.2.

The proof of Lemma 4.3 is presented in Appendix C.2. Combining Lemma 4.3 with Lemma 4.1, we obtain a direct bound of the Δ_{DP} in terms of the training regularizer U_{CHSW}.

Theorem 4.4 (Upper bounds for Δ_{DP} on pullback Euclidean manifolds)**.** *Under the assumptions of Lemma 4.3, we have*

$$\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta) \ \le\ 2\sqrt{2\rho\, L}\, C_{d,p,r}^{\frac{1}{2p}} \left(\frac{2^p}{\pi_{\min}} U_{\mathrm{CHSW}}\right)^{\frac{1}{2(d+1)p^2}} \tag{7}$$

Substituting Equation (4) into Equation (6) proves Equation (7). Therefore, under the above assumptions on h, $\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta)$ is bounded by $U_{\mathrm{CHSW}}^{\frac{1}{2(d+1)p^2}}$, providing a theoretical justification for adopting U_{CHSW} as a fairness regularizer.

4.3 All Hadamard Manifolds

On any Hadamard manifold, CHSW_p is a finite pseudo-distance on $\mathcal{P}_p(\mathcal{M})$ (Bonet et al., 2025, Proposition 17): it is nonnegative, symmetric, satisfies the triangle inequality, and is finite for all $\mu, \nu \in \mathcal{P}_p(\mathcal{M})$. However, the identity of indiscernibles (i.e., $\mathrm{CHSW}_p(\mu,\nu) = 0 \Rightarrow \mu = \nu$) remains unestablished in full generality. Moreover, Proposition 4.2 no longer applies beyond the pullback Euclidean setting. As a result, we cannot directly induce a strict implication between the Δ_{DP} and our CHSW-based $\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}$ (and hence U_{CHSW}). We therefore use a soft demographic parity violation for general Hadamard manifolds, while reporting Δ_{DP} empirically.

We consider a score head $h : \mathcal{M} \to [0,1]$ operating on a representation $Z \in \mathcal{M}$. Given group-conditional distributions $\mu_s := (f_\theta)_\# P_{X|S=s}$, we consider the soft demographic parity (Agarwal et al., 2018; Wei et al., 2023):

$$\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h \circ f_\theta) \;=\; \max_{s,s' \in [A]} |\mathbb{E}[h(Z) \mid S = s] - \mathbb{E}[h(Z) \mid S = s']| = \max_{s,s' \in [A]} \left| \int_{\mathcal{M}} h \, d\mu_s - \int_{\mathcal{M}} h \, d\mu_{s'} \right|.$$

Theorem 4.5 (Upper bounds for $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$ on Hadamard manifolds)**.** *Let $(\mathcal{M}, d)$ be a Hadamard manifold with origin o, and let λ_o be the uniform distribution over $S_o = \{v \in T_o\mathcal{M} : \|v\|_o = 1\}$. Fix $p \geq 1$ and consider two group-conditional distributions $\mu_s, \mu_{s'} \in \mathcal{P}_p(\mathcal{M})$. Assume that, for λ_o-a.e. $v \in S_o$, the projection $P_v : \mathcal{M} \to \mathbb{R}$ is measurable, normalized as $P_v(o) = 0$, and 1-Lipschitz: $|P_v(z) - P_v(z')| \leq d(z, z'), \forall z, z' \in \mathcal{M}$. Assume further that the score head $h : \mathcal{M} \to \mathbb{R}$ admits the sliced representation*

$$h(z) = \int_{S_o} g_v(P_v(z)) \, d\lambda_o(v), \tag{8}$$

where $(v, t) \mapsto g_v(t)$ is measurable and satisfies the following uniform regularity conditions:

*1. **(Uniform Lipschitzness)** $\mathrm{Lip}(g_v) \leq L_{\mathrm{slice}}$ for λ_o-a.e. v;*

*2. **(Uniform linear growth)** There exists $C_0 \geq 0$ such that for λ_o-a.e. v and all $t \in \mathbb{R}$,$|g_v(t)| \leq C_0 + L_{\mathrm{slice}}|t|$. Then,*

$$\left| \int_{\mathcal{M}} h \, d\mu_s - \int_{\mathcal{M}} h \, d\mu_{s'} \right| \leq L_{\mathrm{slice}} \cdot \Big(\mathrm{CHSW}_p^p(\mu_s, \mu_{s'}) \Big)^{1/p}. \tag{9}$$

Consequently, the soft demographic parity violation obeys

$$\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h \circ f_\theta) \;\leq\; 2 \, L_{\mathrm{slice}} \, U_{\mathrm{CHSW}}^{1/p} \, \pi_{\min}^{-1/p}. \tag{10}$$

The proof of Theorem 4.5 is presented in Appendix C.4. Therefore, under the above assumptions on h, $\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h \circ f_\theta)$ is controlled by $U_{\mathrm{CHSW}}^{1/p}$, which justifies using U_{CHSW} as a theoretically grounded fairness regularizer. The assumption in Theorem 4.5 is analytical and is discussed in Appendix C.5.

4.4 Training Objective and Practical Computation

Therefore, the loss function is formed as

$$\mathcal{L}(\theta, \psi) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathrm{BCE}}(h_\psi(f_\theta(X_i)) \, , \, Y_i) + \lambda \cdot U_{\mathrm{CHSW}}(\{Z_i\}_{i=1}^n, \{S_i\}_{i=1}^n),$$

where $\lambda \geq 0$ controls the fairness–accuracy trade-off. In practice, we approximate the integral over S_o using K random directions $\{v_k\}_{k=1}^K$:

$$\mathrm{CHSW}_p^p(\mu, \nu) \approx \frac{1}{K} \sum_{k=1}^{K} W_p^p\big((\Pi_{v_k})_\# \mu, \; (\Pi_{v_k})_\# \nu\big),$$

where Π_v denotes the 1D projection along direction v, given by the geodesic projection P_v or the horospherical projection B_v. Each 1D Wasserstein distance is computed via quantiles using a closed-form expression on $\mathbb{R}$ (Bonet et al., 2025). In practice, Algorithm 1 optimizes a tractable surrogate $\widetilde{U}_{\mathrm{CHSW}}$ of U_{CHSW} by decoupling barycenters across slices, while preserving the pairwise CHSW control needed for our fairness guarantees. More details, pseudocode, and complexity are provided in Appendix D.

5 Experiments

In this section, we evaluate our proposed method on both synthetic and real-world datasets to assess its ability to simultaneously control fairness and predictive accuracy.

5.1 Experimental Details

5.1.1 Baselines and metrics.

We compare against several regularization-based FRL baselines, which enforce fairness by pulling different groups' representations close under various probability metrics: (1) $\mathbf{M_{AVG}}$ (Oneto et al., 2020) aligns sensitive-group representations by *mean matching*, which is equivalent to MMD with a linear kernel. (2) $\mathbf{M_{MMD}}$ (Oneto et al., 2020) enforces demographic parity by minimizing MMD with a *Gaussian kernel*, capturing richer distribution discrepancies beyond first-order moments. (3) $\mathbf{M_{SNK}}$ (Oneto et al., 2020) matches sensitive-group distributions using the *Sinkhorn divergence*, an entropic optimal-transport based discrepancy for representation alignment. (4) **WB** (Gaucher et al., 2023) constructs a fair predictor by transporting group-wise score distributions toward their *Wasserstein-2 barycenter* and then thresholding the resulting fair regressor. (5) **SW** (Nguyen et al., 2025) computes a geometry-aware sliced Wasserstein discrepancy in Euclidean latent space via random projections, serving as an efficient fairness baseline. To keep the training protocol consistent across baselines, we adapt SW accordingly. For the comparison across different methods, we use accuracy (**ACC**) to measure the predictive performance. To assess fairness, we measure the degree of fairness violations using both the standard demographic parity $\mathbf{\Delta_{DP}}$ and its soft relaxation $\mathbf{\Delta_{DP}^{soft}}$.

5.1.2 Datasets and implementation.

We validate the efficacy of our proposed method on three synthetic datasets defined on different Hadamard manifolds and two real-world benchmarks. (1) Pullback Euclidean Space Synthetic Datasets (**Pullback**): we construct a binary classification dataset with group-dependent base-rate bias and nonlinear features that become Euclidean under the pullback transform $\phi(\cdot) = \operatorname{asinh}(\cdot)$. (2) Hyperbolic Space Synthetic Datasets (**Hyperbolic**): we synthesize data on the Poincaré ball (curvature -1) with two group centers separated by a fixed hyperbolic distance, inducing an intrinsic hyperbolic distribution shift. (3) SPD Space Synthetic Datasets (**SPD**): we generate representations on the SPD manifold, enabling evaluation under AIRM-induced non-Euclidean geometry. (4) Adult Census (**Adult**) Dataset (Kohavi, 1996): we use the Adult Income dataset to predict whether an individual earns more than \$50K/year, with gender as the sensitive attribute. (5) Labeled Faces in the Wild with Attributes Dataset (**LFWA**) (Kumar et al., 2009; Wolf et al., 2010): we use the LFWA benchmark, where each face image is augmented with attributes such as gender and race, and the task is to predict gender while treating "WavyHair" as the sensitive attribute, following prior work that reports its strong association with male labels. Since the intrinsic geometry of real-world representation spaces is unknown, we evaluate several candidate manifolds and select the best per dataset (e.g., pullback Euclidean for Adult, hyperbolic for LFWA), following standard practice in geometry-aware representation learning (Bachmann et al., 2020; Chami et al., 2019; Gu et al., 2018). For all baselines we use the original papers' hyperparameters, set $p = 2$ for CHSW_p^p, and evaluate all methods under a consistent protocol. Details are in Appendix E.

5.2 Main Results

5.2.1 Fairness performance.

To evaluate fairness efficiency, we compare against five state-of-the-art baselines on simulation data from three Hadamard manifold settings and two real-world datasets as shown in Table 1. Our method achieves consistently lower Δ_{DP} and $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$ across almost all settings while keeping ACC competitive. By contrast, $\mathrm{M_{AVG}}$, $\mathrm{M_{MMD}}$, and $\mathrm{M_{SNK}}$ tend to exhibit lower ACC and higher Δ_{DP} and $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$ on most datasets. We conjecture that this is mainly because these methods measure distributional discrepancies, and hence assess fairness, primarily in Euclidean space. When the underlying representation space exhibits pronounced non-Euclidean geometry, Euclidean metrics may fail to faithfully capture the true differences between distributions, leading to biased fairness estimation. WB and SW fare better on the hyperbolic and SPD settings and on Adult and LFWA, as both exploit geometric distributional cues; but WB uses the Wasserstein barycenter of score distributions and SW the sliced-Wasserstein barycenter of representations, neither of which captures manifold-level discrepancies, so both still trail our method. Overall, measuring distributional distances under geometry-consistent manifolds yields a more faithful fairness assessment, giving higher accuracy and smaller fairness gaps across diverse settings.

Dataset	Metric	Model					
		M_{AVG}	M_{MMD}	M_{SNK}	WB	SW	Ours
Pullback	ACC	0.671 ± 0.026	0.754 ± 0.045	0.754 ± 0.178	0.706 ± 0.041	0.724 ± 0.041	**0.758 ± 0.064**
	Δ_{DP}	0.214 ± 0.078	0.192 ± 0.077	0.389 ± 0.358	0.212 ± 0.090	0.244 ± 0.087	**0.187 ± 0.122**
	Δ_{DP}^{soft}	0.179 ± 0.047	0.185 ± 0.059	0.644 ± 0.075	0.476 ± 0.113	0.209 ± 0.073	**0.176 ± 0.096**
Hyperbolic	ACC	0.517 ± 0.089	0.462 ± 0.075	0.496 ± 0.064	0.518 ± 0.013	0.519 ± 0.064	**0.521 ± 0.061**
	Δ_{DP}	0.379 ± 0.224	0.393 ± 0.154	0.270 ± 0.186	**0.169 ± 0.075**	0.288 ± 0.154	0.245 ± 0.158
	Δ_{DP}^{soft}	0.067 ± 0.068	0.029 ± 0.019	0.009 ± 0.004	0.026 ± 0.006	0.010 ± 0.008	**0.005 ± 0.006**
SPD	ACC	0.784 ± 0.025	0.791 ± 0.011	0.788 ± 0.016	0.793 ± 0.018	0.794 ± 0.024	**0.795 ± 0.034**
	Δ_{DP}	0.215 ± 0.090	0.224 ± 0.021	0.223 ± 0.083	0.247 ± 0.050	0.232 ± 0.077	**0.210 ± 0.145**
	Δ_{DP}^{soft}	0.101 ± 0.043	0.122 ± 0.019	0.104 ± 0.036	0.111 ± 0.008	0.110 ± 0.037	**0.079 ± 0.068**
Adult	ACC	0.837 ± 0.016	0.836 ± 0.002	0.839 ± 0.006	0.831 ± 0.005	0.825 ± 0.040	**0.840 ± 0.007**
	Δ_{DP}	0.112 ± 0.065	0.113 ± 0.030	0.107 ± 0.020	0.107 ± 0.008	0.102 ± 0.059	**0.098 ± 0.034**
	Δ_{DP}^{soft}	0.087 ± 0.040	0.087 ± 0.022	0.085 ± 0.017	0.116 ± 0.005	0.085 ± 0.048	**0.082 ± 0.033**
LFWA	ACC	0.900 ± 0.005	0.902 ± 0.004	0.894 ± 0.009	0.902 ± 0.005	0.899 ± 0.010	**0.905 ± 0.004**
	Δ_{DP}	0.056 ± 0.013	0.056 ± 0.008	0.056 ± 0.013	0.058 ± 0.011	0.051 ± 0.015	**0.048 ± 0.012**
	Δ_{DP}^{soft}	0.051 ± 0.013	0.049 ± 0.007	0.050 ± 0.015	0.046 ± 0.007	0.048 ± 0.012	**0.046 ± 0.010**

Table 1: Performance comparison in terms of ACC, Δ_{DP} and Δ_{DP}^{soft} on five datasets (Format: mean±std). The best results are highlighted in **bold**, while the second-best results are underlined.

5.2.2 Efficiency performance.

In terms of efficiency, as shown in Table 2, CHSW is not the fastest method due to the sorting operation required in sliced Wasserstein computation. Nevertheless, its runtime (RT) and memory (Mem) overhead remain comparable to other FRL baselines, such as M_{MMD} and SW, and are substantially lower than M_{SNK}. These results indicate that the computational cost of our method remains practical in both synthetic and real-world settings.

Table 2: Runtime (ms/epoch) and memory (MB). Runtime is averaged over 10 epochs; WB (post-processing) reports one-time cost. Memory denotes peak GPU usage during training.

	Pullback		Hyperbolic		SPD		Adult		LFWA	
Method	RT	Mem	RT	Mem	RT	Mem	RT	Mem	RT	Mem
M_{AVG}	41.9	0.8	240.0	0.7	17.1	18.6	765.3	1.6	1680.3	188.6
M_{MMD}	36.3	2.3	256.2	2.2	21.9	20.2	990.6	3.2	1673.7	188.6
M_{SNK}	420.7	2.3	3098.0	2.3	255.9	20.2	18115.7	22.9	4058.8	188.6
WB	25.6	0.8	160.8	0.7	18.2	18.6	483.7	1.6	1828.7	358.5
SW	43.0	2.2	297.9	2.2	26.2	19.2	920.3	3.1	2187.1	358.5
Ours	48.9	2.3	321.9	6.9	36.0	19.3	925.5	3.1	2095.7	188.6

5.3 Representation & Prediction Distributions Analysis

We visualize PCA scatter plots of the learned representations Z and kernel density estimates (KDEs) of prediction scores $\hat{Y}$, both stratified by sensitive groups, to examine group-level distributions in the representation and prediction spaces. Figure 2 reports results on Adult using gender as the sensitive attribute, while Figure 5 (Appendix E.4) reports results on LFWA using WavyHair as the sensitive attribute. The top row shows the learned representations Z of test samples. Since Z is 16-dimensional, we project it onto two dimensions using PCA, where the x- and y-axes correspond to PC1 and PC2. The bottom row reports the KDEs of prediction scores, also stratified by sensitive groups. We observe that, at comparable accuracy, our method produces better group alignment in Z, which further leads to greater overlap in the score distributions, underscoring the effectiveness of our approach.

5.4 Ablation Studies

Regularization coefficient λ. We study the fairness–accuracy trade-off on synthetic data by sweeping $\lambda \in \{0, 10^{-2}, 10^{-1}, \ldots, 10^{14}\}$. As shown in Figure 3, our method consistently achieves lower Δ_{DP} and Δ_{DP}^{soft} at comparable accuracy, and higher accuracy under the same fairness level. This demonstrates improved trade-off behavior across a wide range of λ, which we attribute to its ability to capture distributional fairness directly on the manifold.

Representation dimension d_z and hidden dimension h. We analyze the impact of the representation dimension d_z (16 – 128) and hidden dimension h (16 – 128) on LFWA. As shown in Figure 6 and Figure 7 (Appendix E.5), our method is stable across a wide range of settings, with performance degradation only at extreme configurations ($d_z \in \{16, 128\}$ or $h = 16$). We attribute this to limited

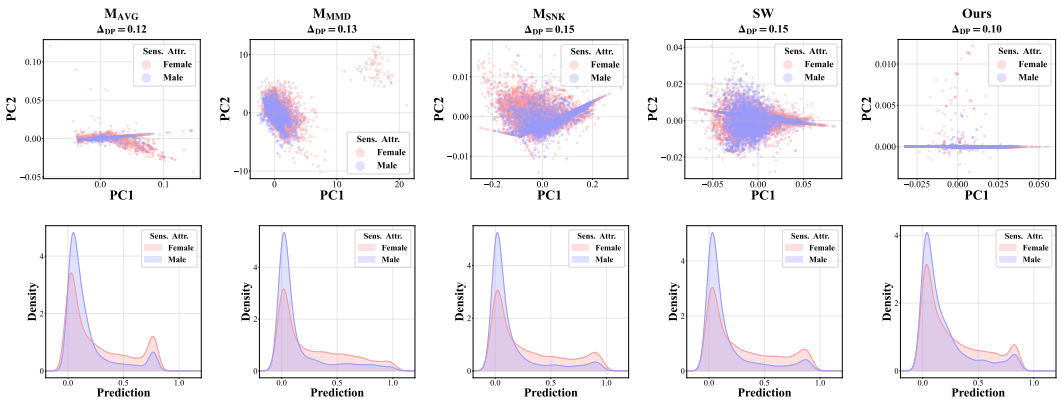
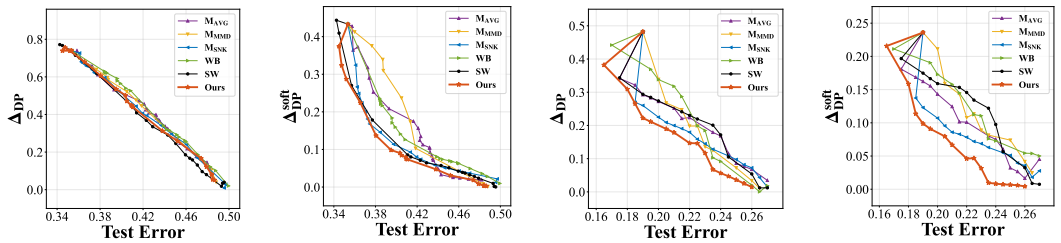


Figure 2: Representation (top) and prediction distributions (bottom) analysis on the Adult dataset.



(a) Δ_{DP} on Hyperbolic (b) $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$ on Hyperbolic (c) Δ_{DP} on SPD (d) $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$ on SPD

Figure 3: The fairness–accuracy trade-off of the trained models on Hyperbolic and SPD datasets.

effective capacity: small d_z or h restricts modeling flexibility, while overly large d_z may lead to overfitting and reduced fairness.

Number of CHSW projection directions K. We further evaluate the sensitivity to the number of projection directions K, as summarized in Table 4 in Appendix E.5. Overall, the results are stable across different values of K, with moderate values (e.g., 120) achieving a good balance between performance and efficiency. This demonstrates that our method is not overly sensitive to these hyperparameters and can be applied with simple default choices.

Origin choice o in CHSW. Our implementation follows the standard CHSW formulation with a fixed reference origin (e.g., ball center or identity)(Bonet et al., 2025). We further evaluate multiple alternatives, including random and data-adaptive choices. As shown in Table 3 in Appendix E.5, although different origins lead to some variation, the overall performance remains stable, and our method consistently improves fairness metrics. Therefore, CHSW is robust to the choice of origin in practice, and simple canonical choices already work well.

6 Conclusions

This paper proposes a manifold-based FRL framework that relaxes Euclidean assumptions in prior work and incorporates manifold geometry for real-world data distributions. To quantify distributional discrepancies as a fairness criterion in such settings, we introduce CHSW, a computationally efficient discrepancy that captures differences between distributions on Hadamard manifolds. We further provide theoretical guarantees showing that minimizing our CHSW-based regularization provably controls demographic parity violation with explicit bounds. Extensive experiments on standard benchmarks demonstrate that our method consistently outperforms state-of-the-art baselines, achieving competitive accuracy while substantially improving fairness. These contributions advance geometry-consistent fairness learning beyond Euclidean latent spaces and lay a principled foundation for future manifold-aware, reliable, and scalable FRL.

AI USE STATEMENT

AI-based tools were used during the preparation of this manuscript for limited purposes, including language editing, proofreading, and assisting with code implementation. All mathematical derivations, theoretical results, algorithmic designs, and scientific conclusions were developed by the authors. All AI-assisted code was manually reviewed, verified, and validated by the authors, and all experimental implementations and empirical analyses were independently checked by the authors.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *International conference on machine learning*, pp. 486–496. PMLR, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

AR Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 334–370. PMLR, 2023a.

Clément Bonet, Benoıt Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-wasserstein on symmetric positive definite matrices for m/eeg signals. In *International Conference on Machine Learning*, pp. 2777–2805. PMLR, 2023b.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-wasserstein distances and flows on cartan-hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Shubham Chowdhary, Giulia De Pasquale, Nicolas Lanzetti, Ana-Andreea Stoica, and Florian Dorfler. Fairness in social influence maximization via optimal transport. *Advances in Neural Information Processing Systems*, 37:10380–10413, 2024.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331, 2020.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445. PMLR, 2019.

Friso de Kruiff, Erik J Bekkers, Ozan Öktem, Carola-Bibiane Schönlieb, and Willem Diepeveen. Pullback flow matching on data manifolds. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

David Quashigah Dzakpasu, Jixue Liu, Jiuyong Li, and Lin Liu. Integrating fair representation learning with fairness regularization for intersectional group fairness. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 560–569, 2024.

Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models, 2021. URL https://arxiv.org/abs/1906.08039.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with wasserstein barycenters for non-decomposable performance measures. In *International Conference on Artificial Intelligence and Statistics*, pp. 2436–2459. PMLR, 2023.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365. PMLR, 2019.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2018.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 202–207. AAAI Press, 1996.

Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3): 1083–1094, 2022.

Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 365–372, 2009. doi: 10.1109/ICCV.2009.5459250.

Serge Lang. *Fundamentals of differential geometry*, volume 191. Springer Science & Business Media, 2012.

John M Lee. *Riemannian manifolds: An introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, 2022.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder, 2017. URL https://arxiv.org/abs/1511.00830.

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

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8 (1):141–163, 2021.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31, 2018.

Khai Nguyen, Hai Nguyen, and Nhat Ho. Towards marginal fairness sliced wasserstein barycenter. In *International Conference on Learning Representations*, volume 2025, pp. 54466–54494, 2025.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33:15360–15370, 2020.

Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

Joel W Robbin and Dietmar A Salamon. *Introduction to differential geometry*. Springer Nature, 2022.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Raif M Rustamov and Subhabrata Majumdar. Intrinsic sliced wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, pp. 29388–29415. PMLR, 2023.

Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2020.

Sunzida Siddique, Mohd Ariful Haque, Roy George, Kishor Datta Gupta, Debashis Gupta, and Md Jobair Hossain Faruk. Survey on machine learning biases and mitigation techniques. *Digital*, 4(1):1–68, 2023.

Alberto Sinigaglia, Davide Sartor, Marina Ceccon, and Gian Antonio Susto. Simple and effective specialized representations for fair classifiers. *Advances in Neural Information Processing Systems*, 38:18058–18085, 2026.

Yijie Wang, Viet Anh Nguyen, and Grani A Hanasusanto. Wasserstein robust classification with fairness constraints. *Manufacturing & Service Operations Management*, 26(4):1567–1585, 2024.

Shaokui Wei, Jiayin Liu, Bing Li, and Hongyuan Zha. Mean parity fair regression in rkhs. In *International Conference on Artificial Intelligence and Statistics*, pp. 4602–4628. PMLR, 2023.

Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1978–1990, 2010.

Zikai Xiong, Niccolò Dalmasso, Alan Mishler, Vamsi K Potluru, Tucker Balch, and Manuela Veloso. Fairwasp: Fast and optimal fair wasserstein pre-processing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):16120–16128, 2024a.

Zikai Xiong, Niccolò Dalmasso, Shubham Sharma, Freddy Lecue, Daniele Magazzeni, Vamsi Potluru, Tucker Balch, and Manuela Veloso. Fair wasserstein coresets. *Advances in Neural Information Processing Systems*, 37:132–168, 2024b.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3278721.3278779.

A MOTIVATING ILLUSTRATION FOR MANIFOLD-AWARE FAIRNESS

As shown in Figure 1 and Figure 4, distributions that look nearly identical under Euclidean measures can be quite different on the true underlying manifold. This mismatch means that representations deemed "fair" in Euclidean space may actually preserve group-specific patterns when viewed through the correct geometric lens, rendering existing fairness guarantees unreliable.

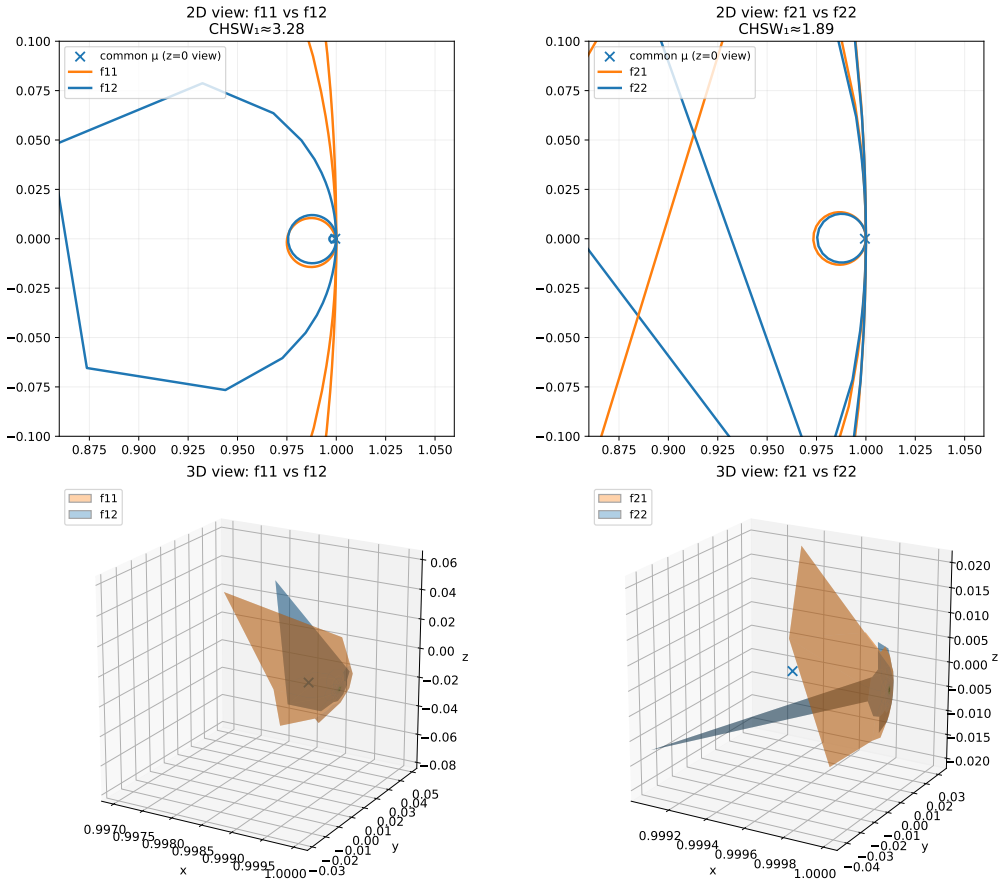


Figure 4: Illustration that equal Euclidean parameter distances do not imply similar distributions on a manifold. We form two pairs of wrapped normal distributions on the 3D Poincaré ball $\mathbb{B}^3$ ($K = -1$) with a shared mean $\mu = (0.9995, 0, 0)$: $f_{ij} = (\exp_\mu)_\# \mathcal{N}(0, \Sigma_{ij})$ for $(i, j) \in \{1, 2\}^2$ (covariances $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ are given in Appendix A). The Frobenius (Euclidean) parameter gaps are matched, $d_F(\Sigma_a, \Sigma_b) = \|\Sigma_a - \Sigma_b\|_F$ and $\|\Sigma_{11} - \Sigma_{12}\|_F = \|\Sigma_{21} - \Sigma_{22}\|_F \approx 2.57 \times 10^{-5}$, yet the manifold distribution distances differ markedly under horospherical CHSW_1, $\mathrm{CHSW}_1(f, g) = \mathbb{E}_{v \sim \mathrm{Unif}(\mathbb{S}^2)} W_1((b_v)_\# f, (b_v)_\# g)$ with $b_v(x) = \log(\|v - x\|^2/(1 - \|x\|^2))$: $\mathrm{CHSW}_1(f_{11}, f_{12}) \approx 3.28$ vs. $\mathrm{CHSW}_1(f_{21}, f_{22}) \approx 1.89$ (gap ≈ 1.39). This shows that matching Euclidean parameter distances does not imply similar manifold distributions, motivating distribution distances defined directly on the manifold.

A.1 WRAPPED NORMAL DISTRIBUTIONS ON THE POINCARÉ BALL

All distributions in Figure 1 and Figure 4 are wrapped normal distributions on the 3D Poincaré ball $\mathbb{B}^3 = \{x \in \mathbb{R}^3 : \|x\| < 1\}$ with constant curvature $K = -1$. Given a mean $\mu \in \mathbb{B}^3$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{3\times 3}$, the wrapped normal distribution is defined as the pushforward

$$f = (\exp_\mu)_\# \mathcal{N}(0, \Sigma),$$

where $\exp_\mu : T_\mu\mathbb{B}^3 \to \mathbb{B}^3$ denotes the Riemannian exponential map under the Poincaré metric. In all experiments, the common mean is fixed to

$$\mu = (0.9995,\, 0,\, 0),$$

placing the distributions close to the boundary to emphasize curvature effects.

A.2 Covariance Parameter Construction

We construct two pairs of covariance matrices $(\Sigma_{11}, \Sigma_{12})$ and $(\Sigma_{21}, \Sigma_{22})$, all symmetric positive definite. The explicit matrices are

$$\Sigma_{11} = 10^{-6}\begin{pmatrix} 30 & 6 & 4 \\ 6 & 6 & 1 \\ 4 & 1 & 6 \end{pmatrix}, \qquad \Sigma_{12} = 10^{-6}\begin{pmatrix} 6 & 0.6 & 0.4 \\ 0.6 & 6 & 1 \\ 0.4 & 1 & 6 \end{pmatrix},$$

and

$$\Sigma_{21} = 10^{-6}\begin{pmatrix} 25 & -2 & 1 \\ -2 & 25 & -1 \\ 1 & -1 & 20 \end{pmatrix}, \quad \Sigma_{22} = \Sigma_{21} + Q(\Sigma_{11} - \Sigma_{12})Q^\top, \quad Q = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

By construction, the Frobenius distances between the two covariance pairs are identical:

$$\|\Sigma_{11} - \Sigma_{12}\|_F = \|\Sigma_{21} - \Sigma_{22}\|_F \approx 2.57 \times 10^{-5}.$$

All covariance matrices have strictly positive minimum eigenvalues, ensuring positive definiteness.

B Examples of Hadamard Manifolds

This section provides concrete examples of Hadamard manifolds used or discussed in this work, including pullback Euclidean spaces, hyperbolic spaces, and the manifold of symmetric positive-definite matrices.

B.1 Pullback Euclidean Spaces

Definition B.1 (Pullback Euclidean Space)**.** Let $(\mathcal{N}, \langle\cdot,\cdot\rangle)$ be a Euclidean space and let $\phi : \mathcal{M} \to \mathcal{N}$ be a diffeomorphism. The pullback Euclidean metric induced by ϕ is defined, for any $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M}$, by

$$g_x^\phi(u, v) = \big\langle \phi_{*,x}(u), \phi_{*,x}(v) \big\rangle,$$

where $\phi_{*,x}$ denotes the differential of ϕ at x, and $T_x\mathcal{M}$ is the tangent space of $\mathcal{M}$ at x. Under this metric, the geodesic distance has the closed form

$$d_{\mathcal{M}}(x, y) = \|\phi(x) - \phi(y)\|,$$

and the exponential map is given by

$$\exp_x(v) = \phi^{-1}\big(\phi(x) + \phi_{*,x}(v)\big).$$

Therefore, pullback Euclidean manifolds are flat, with zero sectional curvature, and form a basic subclass of Hadamard manifolds. Euclidean space is recovered as the special case where ϕ is the identity map.

B.2 Hyperbolic Spaces

Definition B.2 (Hyperbolic Space)**.** Hyperbolic space is a complete, simply connected Riemannian manifold with constant negative sectional curvature $K < 0$, and hence is a Hadamard manifold.

A common realization is the Lorentz model:

$$\mathbb{L}_K^d = \Big\{ x = (x_0, \ldots, x_d) \in \mathbb{R}^{d+1} : \langle x, x\rangle_L = \frac{1}{K},\ x_0 > 0 \Big\},$$

where the Minkowski bilinear form is defined as

$$\langle x, y\rangle_L = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i.$$

The geodesic distance in this model is

$$d_{\mathbb{L}}(x, y) = \frac{1}{\sqrt{-K}} \operatorname{arccosh}\big(K\langle x, y\rangle_L\big).$$

B.3 Manifold of Symmetric Positive-Definite Matrices

Definition B.3 (Manifold of Symmetric Positive-Definite Matrices)**.** Let $\mathbb{S}^d(\mathbb{R})$ denote the space of real symmetric $d \times d$ matrices. The manifold of symmetric positive-definite matrices is

$$\mathbb{S}^d_{++}(\mathbb{R}) = \Big\{ M \in \mathbb{S}^d(\mathbb{R}) : x^\top M x > 0,\ \forall x \in \mathbb{R}^d \setminus \{0\} \Big\}.$$

It is a smooth manifold, and the tangent space at M can be identified with $\mathbb{S}^d(\mathbb{R})$.

A classical metric on this manifold is the affine-invariant Riemannian metric (AIRM). For $B_1, B_2 \in T_M \mathbb{S}^d_{++}(\mathbb{R})$, it is defined as

$$\langle B_1, B_2 \rangle_M = \mathrm{Tr}\big(M^{-1} B_1 M^{-1} B_2\big).$$

The corresponding geodesic distance is

$$d_{\mathrm{AI}}(X, Y) = \Big\| \log\Big(X^{-1/2} Y X^{-1/2} \Big) \Big\|_F .$$

Under the affine-invariant Riemannian metric, $\mathbb{S}^d_{++}(\mathbb{R})$ is complete and simply connected with non-positive sectional curvature, hence it is a Hadamard manifold.

C Methodology Details

C.1 Proof of Lemma 4.1

Proof. Recall that the group-wise representation distributions are denoted by $\mu_s \triangleq (f_\theta)_\# P_{X|S=s}$ for $s \in [A]$, and the distributional DP deviation is

$$\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta) = \max_{s,t \in [A]} \mathrm{CHSW}_p^p(\mu_s, \mu_t).$$

Let $\bar{\mu}$ be the reference distribution used in the regularizer

$$U_{\mathrm{CHSW}} \triangleq \sum_{s=0}^{A-1} \pi_s \, \mathrm{CHSW}_p^p(\mu_s, \bar{\mu}),$$
$$\text{with } \ \pi_s > 0, \ \ \sum_{s=0}^{A-1} \pi_s = 1, \ \ \pi_{\min} = \min_s \pi_s > 0.$$

Fix any $s, t \in [A]$. Since CHSW_p is a metric, it satisfies the triangle inequality:

$$\mathrm{CHSW}_p(\mu_s, \mu_t) \le \mathrm{CHSW}_p(\mu_s, \bar{\mu}) + \mathrm{CHSW}_p(\mu_t, \bar{\mu}). \tag{11}$$

On the other hand, by the definition of U_{CHSW},

$$U_{\mathrm{CHSW}} = \sum_{r=0}^{A-1} \pi_r \, \mathrm{CHSW}_p^p(\mu_r, \bar{\mu}) \ge \pi_s \, \mathrm{CHSW}_p^p(\mu_s, \bar{\mu}),$$

which implies

$$\mathrm{CHSW}_p(\mu_s, \bar{\mu}) \le \left(\frac{U_{\mathrm{CHSW}}}{\pi_s} \right)^{1/p}. \tag{12}$$

Similarly,

$$\mathrm{CHSW}_p(\mu_t, \bar{\mu}) \le \left(\frac{U_{\mathrm{CHSW}}}{\pi_t} \right)^{1/p}. \tag{13}$$

Substituting Equation (12)–Equation (13) into Equation (11) yields

$$\mathrm{CHSW}_p(\mu_s, \mu_t) \le \left(\frac{U_{\mathrm{CHSW}}}{\pi_s} \right)^{1/p} + \left(\frac{U_{\mathrm{CHSW}}}{\pi_t} \right)^{1/p}.$$

Taking the p-th power on both sides, and applying the inequality $(a+b)^p \le 2^{p-1}(a^p+b^p)$ for $a, b \ge 0$, we obtain

$$\begin{aligned}\mathrm{CHSW}_p^p(\mu_s,\mu_t) &\le \left(\left(\frac{U_{\mathrm{CHSW}}}{\pi_s}\right)^{1/p} + \left(\frac{U_{\mathrm{CHSW}}}{\pi_t}\right)^{1/p}\right)^p \\ &\le 2^{p-1}\left(\frac{U_{\mathrm{CHSW}}}{\pi_s} + \frac{U_{\mathrm{CHSW}}}{\pi_t}\right) \\ &= 2^{p-1} U_{\mathrm{CHSW}}\left(\frac{1}{\pi_s} + \frac{1}{\pi_t}\right). \end{aligned} \tag{14}$$

Finally, since $\pi_s, \pi_t \ge \pi_{\min}$, we have

$$\frac{1}{\pi_s} + \frac{1}{\pi_t} \le \frac{2}{\pi_{\min}},$$

and thus Equation (14) implies

$$\mathrm{CHSW}_p^p(\mu_s,\mu_t) \le \frac{2^p}{\pi_{\min}} U_{\mathrm{CHSW}}.$$

Taking the maximum over all $s, t \in [A]$, we conclude that

$$\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta) = \max_{s,t\in[A]} \mathrm{CHSW}_p^p(\mu_s,\mu_t) \le \frac{2^p}{\pi_{\min}} U_{\mathrm{CHSW}},$$

which completes the proof. □

C.2 Proof of Lemma 4.3

Proof. Fix any $s, t \in [A]$. Denote $U_s \triangleq h(Z) \mid S = s$ and $U_t \triangleq h(Z) \mid S = t$. Let $\tau \in (0,1)$ be the fixed threshold and take any $\epsilon > 0$. Define the ϵ-smoothed step function $\psi_\epsilon : [0,1] \to [0,1]$ by the clipped ramp

$$\begin{aligned}\psi_\epsilon(u) &= \left[\frac{u-(\tau-\epsilon)}{2\epsilon}\right]_{[0,1]} \\ &= \min\left\{1, \max\left\{0, \frac{u-\tau+\epsilon}{2\epsilon}\right\}\right\}, \qquad u \in [0,1].\end{aligned}$$

This clipped definition remains valid even when $\tau - \epsilon < 0$ or $\tau + \epsilon > 1$. Since the clipping map is 1-Lipschitz and the affine map has slope $1/(2\epsilon)$, ψ_ϵ is $1/(2\epsilon)$-Lipschitz on $[0,1]$. Moreover, for any $u \in [0,1]$,

$$\begin{aligned}\mathbf{1}\{u \ge \tau\} &\le \psi_\varepsilon(u) + \mathbf{1}\{|u-\tau| \le \varepsilon\}, \\ \mathbf{1}\{u \ge \tau\} &\ge \psi_\varepsilon(u) - \mathbf{1}\{|u-\tau| \le \varepsilon\}.\end{aligned}$$

Taking expectations with respect to U_s and U_t and subtracting yields

$$\begin{aligned}\Big|\mathbb{P}(U_s \ge \tau) - \mathbb{P}(U_t \ge \tau)\Big| &= \Big|\mathbb{E}\mathbf{1}\{U_s \ge \tau\} - \mathbb{E}\mathbf{1}\{U_t \ge \tau\}\Big| \\ &\le \Big|\mathbb{E}\psi_\varepsilon(U_s) - \mathbb{E}\psi_\varepsilon(U_t)\Big| + \mathbb{P}(|U_s-\tau| \le \varepsilon) + \mathbb{P}(|U_t-\tau| \le \varepsilon). \end{aligned} \tag{15}$$

Bounding the threshold-neighborhood terms. By the local anti-concentration assumption (assumption 2), for all $\varepsilon > 0$,

$$\mathbb{P}(|U_s-\tau| \le \varepsilon) \le 2\rho\varepsilon, \quad \mathbb{P}(|U_t-\tau| \le \varepsilon) \le 2\rho\varepsilon.$$

Therefore, the two neighborhood terms are bounded by $4\rho\varepsilon$.

Bounding the smoothed term via KR duality. Let $\varphi_\varepsilon \triangleq \psi_\varepsilon \circ h : \mathcal{M} \to [0,1]$. By assumption 1 and the Lipschitz property of ψ_ε,

$$\mathrm{Lip}(\varphi_\varepsilon) \le \mathrm{Lip}(\psi_\varepsilon)\,\mathrm{Lip}(h) \le \frac{1}{2\varepsilon}\,L = \frac{L}{2\varepsilon}.$$

By the Kantorovich–Rubinstein duality for W_1,

$$\Big|\mathbb{E}_{Z\sim\mu_s}\varphi_\varepsilon(Z) - \mathbb{E}_{Z\sim\mu_t}\varphi_\varepsilon(Z)\Big| \le \mathrm{Lip}(\varphi_\varepsilon)\,W_1(\mu_s,\mu_t) \le \frac{L}{2\varepsilon}\,W_1(\mu_s,\mu_t).$$

Since $\mathbb{E}_{Z\sim\mu_s}\varphi_\varepsilon(Z) = \mathbb{E}\psi_\varepsilon(U_s)$ and similarly for t, we obtain

$$\Big|\mathbb{E}\psi_\varepsilon(U_s) - \mathbb{E}\psi_\varepsilon(U_t)\Big| \le \frac{L}{2\varepsilon}\,W_1(\mu_s,\mu_t).$$

Putting the bounds together and optimizing over ε. Substituting the above estimates into Equation (15) gives

$$\Big|\mathbb{P}(U_s \ge \tau) - \mathbb{P}(U_t \ge \tau)\Big| \le \frac{L}{2\varepsilon}\,W_1(\mu_s,\mu_t) + 4\rho\varepsilon.$$

Choosing ε to balance the two terms, e.g.,

$$\varepsilon^\star = \sqrt{\frac{L\,W_1(\mu_s,\mu_t)}{8\rho}},$$

yields

$$\Big|\mathbb{P}(U_s \ge \tau) - \mathbb{P}(U_t \ge \tau)\Big| \le 2\sqrt{2\rho\,L\,W_1(\mu_s,\mu_t)}.$$

Since $W_1 \le (W_p^p)^{1/p}$,

$$\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta) \le 2\sqrt{2\rho\,L}\,\big(W_p^p(\mu_s,\mu_t)\big)^{\frac{1}{2p}}. \tag{16}$$

From W_p^p to CHSW_p^p on pullback Euclidean manifolds. Under assumption 3, each pair (μ_s,μ_t) is supported on $B(o,r)$. On pullback Euclidean manifolds, Bonet et al.($Bonetet\ al.,\ 2025$) proved that there exists $C_{d,p,r} > 0$ such that for any μ,ν supported on $B(o,r)$,

$$W_p^p(\mu,\nu) \le C_{d,p,r}\,\mathrm{CHSW}_p(\mu,\nu)^{\frac{1}{d+1}}.$$

Substituting into Equation (16) yields

$$W_p^p(\mu_s,\mu_t)^{1/2p} \le C_{d,p,r}{}^{\frac{1}{2p}}\,\mathrm{CHSW}_p^p(\mu_s,\mu_t)^{\frac{1}{2(d+1)p^2}}.$$

Therefore, taking the maximum over $s,t \in [A]$ implies

$$\begin{aligned}
&\Delta_{\mathrm{DP}}(c_\tau \circ f_\theta) \\
&\le 2\sqrt{2\rho\,L}\,C_{d,p,r}{}^{\frac{1}{2p}}\,\Big(\max_{s,t\in[A]}\mathrm{CHSW}_p^p(\mu_s,\mu_t)\Big)^{\frac{1}{2(d+1)p^2}} \\
&= 2\sqrt{2\rho\,L}\,C_{d,p,r}{}^{\frac{1}{2p}}\,\Big(\Delta_{\mathrm{DP}}^{\mathrm{CHSW}}(f_\theta)\Big)^{\frac{1}{2(d+1)p^2}}.
\end{aligned}$$

This concludes the proof. □

C.3 Feasibility of Lemma 4.3 Assumptions

Assumption 1 is mild since common score heads (linear + sigmoid/softmax) are Lipschitz on pullback Euclidean manifolds, and the Lipschitz constant can be controlled by standard norm constraints (e.g., spectral normalization).

Assumption 2 is a local anti-concentration condition at the threshold τ. It only restricts how much probability mass the conditional score $U_s \triangleq h(Z) \mid S = s$ can concentrate in an ε-neighborhood of τ: $\mathbb{P}\big(|U_s - \tau| \le \varepsilon\big) \le 2\rho\,\varepsilon, \quad \forall s \in [A],\ \forall \varepsilon > 0$. This condition is mild and feasible in practice. In particular, the condition holds whenever, for each group s, the conditional score $U_s \coloneqq h(Z) \mid S = s$

admits a density p_s that is locally bounded around τ. More precisely, suppose that there exist $\delta > 0$ and $B < \infty$ such that

$$\sup_{s\in[A]} \sup_{u\in(\tau-\delta,\tau+\delta)} p_s(u) \le B.$$

Choose

$$\rho \ge \max\left\{B, \frac{1}{2\delta}\right\}.$$

Then, for any $0 < \epsilon \le \delta$,

$$P(|U_s - \tau| \le \epsilon) = \int_{\tau-\epsilon}^{\tau+\epsilon} p_s(u)\, du \le 2B\epsilon \le 2\rho\epsilon.$$

For $\epsilon > \delta$, the bound follows trivially since

$$P(|U_s - \tau| \le \epsilon) \le 1 \le \frac{\epsilon}{\delta} \le 2\rho\epsilon.$$

Hence the local bounded-density condition implies the anti-concentration assumption for all $\epsilon > 0$ after possibly enlarging ρ.

Assumption 3 can be ensured by projecting representations onto a geodesic ball (equivalently, norm clipping in pullback coordinates), or relaxed to high-probability truncation under mild tail/moment conditions at the cost of an additional tail term.

C.4 Proof of Theorem 4.5

Proof. **Fubini exchange and reduction to one-dimensional pushforwards.** By the sliced representation Equation (8), we have

$$\begin{aligned}&\int_{\mathcal{M}} h\, d\mu_s - \int_{\mathcal{M}} h\, d\mu_{s'} \\ &= \int_{\mathcal{M}} \left(\int_{S_o} g_v(P_v(z))\, d\lambda_o(v)\right) d\mu_s(z) - \int_{\mathcal{M}} \left(\int_{S_o} g_v(P_v(z))\, d\lambda_o(v)\right) d\mu_{s'}(z). \end{aligned} \tag{17}$$

We claim that Fubini's theorem applies so that the order of integration can be exchanged. Indeed, the uniform linear growth condition in Theorem 4.5 implies

$$|g_v(P_v(z))| \le C_0 + L_{\text{slice}}|P_v(z)|.$$

Since $\mu_s, \mu_{s'} \in \mathcal{P}_p(\mathcal{M})$ with $p \ge 1$, in particular $\mu_s, \mu_{s'} \in \mathcal{P}_1(\mathcal{M})$. Moreover, by the projection assumption, for λ_o-a.e. $v \in S_o$, P_v is normalized at the origin and 1-Lipschitz. Hence

$$|P_v(z)| = |P_v(z) - P_v(o)| \le d(z, o), \qquad \forall z \in \mathcal{M},$$

which yields integrability of $|g_v(P_v(z))|$ under μ_s and $\mu_{s'}$ uniformly in v. Therefore, by Fubini's theorem,

$$\int_{\mathcal{M}} h\, d\mu_s - \int_{\mathcal{M}} h\, d\mu_{s'} = \int_{S_o} \left(\int_{\mathcal{M}} g_v(P_v(z))\, d\mu_s(z) - \int_{\mathcal{M}} g_v(P_v(z))\, d\mu_{s'}(z)\right) d\lambda_o(v).$$

Taking absolute values and applying the triangle inequality gives

$$\left|\int_{\mathcal{M}} h\, d\mu_s - \int_{\mathcal{M}} h\, d\mu_{s'}\right| \le \int_{S_o} \left|\int_{\mathcal{M}} g_v(P_v(z))\, d\mu_s(z) - \int_{\mathcal{M}} g_v(P_v(z))\, d\mu_{s'}(z)\right| d\lambda_o(v). \tag{18}$$

Next, by the definition of pushforward measures, for any $\mu \in \mathcal{P}(\mathcal{M})$,

$$\int_{\mathcal{M}} g_v(P_v(z))\, d\mu(z) = \int_{\mathbb{R}} g_v(t)\, d\big((P_v)_{\#}\mu\big)(t). \tag{19}$$

Thus,

$$\left|\int_{\mathcal{M}} g_v(P_v(z))\, d\mu_s(z) - \int_{\mathcal{M}} g_v(P_v(z))\, d\mu_{s'}(z)\right| = \left|\int_{\mathbb{R}} g_v\, d\big((P_v)_{\#}\mu_s\big) - \int_{\mathbb{R}} g_v\, d\big((P_v)_{\#}\mu_{s'}\big)\right|. \tag{20}$$

Kantorovich–Rubinstein duality on $\mathbb{R}$. Since g_v is L_{slice}-Lipschitz for λ_o-a.e. v, the Kantorovich–Rubinstein duality yields for any probability measures α, β on $\mathbb{R}$,

$$\left| \int_{\mathbb{R}} g_v \, d\alpha - \int_{\mathbb{R}} g_v \, d\beta \right| \leq \text{Lip}(g_v) \, W_1(\alpha, \beta) \leq L_{\text{slice}} \, W_1(\alpha, \beta). \tag{21}$$

Applying Equation (21) to $\alpha = (P_v)_{\#}\mu_s$ and $\beta = (P_v)_{\#}\mu_{s'}$, and combining with Equation (18)–Equation (20), we obtain

$$\left| \int_{\mathcal{M}} h \, d\mu_s - \int_{\mathcal{M}} h \, d\mu_{s'} \right| \leq L_{\text{slice}} \int_{S_o} W_1((P_v)_{\#}\mu_s, \, (P_v)_{\#}\mu_{s'}) \, d\lambda_o(v) = L_{\text{slice}} \cdot \text{CHSW}_1(\mu_s, \mu_{s'}). \tag{22}$$

Relating CHSW_1 to CHSW_p. We show that for any $p \geq 1$,

$$\text{CHSW}_1(\mu, \nu) \leq \text{CHSW}_p(\mu, \nu). \tag{23}$$

For each direction $v \in S_o$, the Wasserstein distances on $\mathbb{R}$ satisfy

$$W_1((P_v)_{\#}\mu, (P_v)_{\#}\nu) \leq W_p((P_v)_{\#}\mu, (P_v)_{\#}\nu) \, , \quad (p \geq 1), \tag{24}$$

whenever the p-th moments are finite. Integrating Equation (24) over $v \sim \lambda_o$ yields

$$\text{CHSW}_1(\mu, \nu) = \int_{S_o} W_1((P_v)_{\#}\mu, (P_v)_{\#}\nu) \; d\lambda_o(v) \leq \int_{S_o} W_p((P_v)_{\#}\mu, (P_v)_{\#}\nu) \; d\lambda_o(v).$$

Finally, since λ_o is a probability measure, Hölder's inequality implies

$$\int_{S_o} W_p((P_v)_{\#}\mu, (P_v)_{\#}\nu) \; d\lambda_o(v) \leq \left(\int_{S_o} W_p^p((P_v)_{\#}\mu, (P_v)_{\#}\nu) \; d\lambda_o(v) \right)^{1/p} = \text{CHSW}_p(\mu, \nu),$$

which proves Equation (23). Combining Equation (22) and Equation (23) yields

$$\left| \int_{\mathcal{M}} h \, d\mu_s - \int_{\mathcal{M}} h \, d\mu_{s'} \right| \leq L_{\text{slice}} \cdot \text{CHSW}_p(\mu_s, \mu_{s'}). \tag{25}$$

By definition,

$$\text{CHSW}_p(\mu_s, \mu_{s'}) = \Big(\text{CHSW}_p^p(\mu_s, \mu_{s'}) \Big)^{1/p},$$
$$\text{CHSW}_p^p(\mu, \nu) := \int_{S_o} W_p^p((P_v)_{\#}\mu, (P_v)_{\#}\nu) \; d\lambda_o(v),$$

and therefore Equation (25) is exactly Equation (9).

From pairwise bounds to the soft DP violation. Recall the soft DP violation (over score head h) is

$$\Delta_{\text{DP}}^{\text{soft}}(h \circ f_\theta) := \max_{s,s' \in [A]} \left| \int_{\mathcal{M}} h \, d\mu_s - \int_{\mathcal{M}} h \, d\mu_{s'} \right|.$$

Taking the maximum of Equation (9) over (s, s') yields

$$\Delta_{\text{DP}}^{\text{soft}}(h \circ f_\theta) \leq L_{\text{slice}} \max_{s,s' \in [A]} \Big(\text{CHSW}_p^p(\mu_s, \mu_{s'}) \Big)^{1/p}. \tag{26}$$

Plugging in an upper bound on CHSW_p^p. By Equation (14), established in the proof of Lemma 4.1, for any groups s, t,

$$\text{CHSW}_p^p(\mu_s, \mu_t) \leq 2^{p-1} \, U_{\text{CHSW}} \left(\frac{1}{\pi_s} + \frac{1}{\pi_t} \right). \tag{27}$$

Substituting Equation (27) into Equation (26) gives

$$\begin{aligned} \Delta_{\text{DP}}^{\text{soft}}(h \circ f_\theta) &\leq L_{\text{slice}} \max_{s,t \in [A]} \left(2^{p-1} \, U_{\text{CHSW}} \left(\frac{1}{\pi_s} + \frac{1}{\pi_t} \right) \right)^{1/p} \\ &= L_{\text{slice}} \cdot 2^{\frac{p-1}{p}} \, U_{\text{CHSW}}^{1/p} \cdot \max_{s,t \in [A]} \left(\frac{1}{\pi_s} + \frac{1}{\pi_t} \right)^{1/p}. \end{aligned} \tag{28}$$

Letting $\pi_{\min} := \min_s \pi_s$ and using $\frac{1}{\pi_s} + \frac{1}{\pi_t} \le \frac{2}{\pi_{\min}}$ for all s, t, we obtain

$$\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h \circ f_\theta) \le L_{\mathrm{slice}} \cdot 2^{\frac{p-1}{p}} U_{\mathrm{CHSW}}^{1/p} \cdot \left(\frac{2}{\pi_{\min}}\right)^{1/p}$$
$$= 2\, L_{\mathrm{slice}}\, U_{\mathrm{CHSW}}^{1/p}\, \pi_{\min}^{-1/p},$$

which is exactly Equation (10). This concludes the proof. □

C.5 Feasibility of the Sliced Score-Head Assumption

This section provides additional discussion on the feasibility of the sliced score-head assumption used in Theorem 4.5. Recall that the theorem assumes the downstream score head admits the sliced form

$$h(z) = \int_{S_o} g_v(P_v(z))\, d\lambda(v),$$

where P_v denotes a manifold-aware one-dimensional projection and g_v is a slice-wise scalar function satisfying uniform Lipschitzness and uniform linear growth. This representation is used to connect variations of the score head along one-dimensional projected directions with the CHSW discrepancy between group-conditional representation distributions. Hence, it should be understood as an analytical condition that enables the CHSW-based soft-DP bound, rather than as a restrictive architectural requirement.

C.5.1 A Directly Admissible Finite-Slice Score Head

A simple class of score heads satisfying this condition is the finite-slice logistic head

$$h(z) = \frac{1}{K} \sum_{k=1}^{K} \sigma\big(a_k P_{v_k}(z) + b_k\big), \qquad v_k \sim \lambda_o,$$

where $\sigma(\cdot)$ is the sigmoid function. Here P_{v_k} is chosen according to the underlying geometry, such as geodesic projection in the hyperbolic case, Busemann-type projection in the SPD-AIRM case, and linear projection in the pullback Euclidean case. This finite-slice head can be viewed as a Monte-Carlo approximation of the integral representation, with

$$g_{v_k}(t) = \sigma(a_k t + b_k).$$

The required regularity conditions are mild for this class. Since $\sup_x \sigma'(x) = 1/4$, we have

$$\mathrm{Lip}(g_{v_k}) = \sup_t |g'_{v_k}(t)| \le \frac{|a_k|}{4}.$$

Thus, if $|a_k| \le A$, for example through standard weight decay or weight clipping, then

$$\mathrm{Lip}(g_v) \le L_{\mathrm{slice}} := A/4$$

holds uniformly over the slices. Moreover, since $0 \le \sigma(\cdot) \le 1$, we have

$$|g_{v_k}(t)| \le 1 \le C_0 + L_{\mathrm{slice}}|t|,$$

for instance with $C_0 = 1$ and any $L_{\mathrm{slice}} \ge 0$. Therefore, the uniform linear-growth condition is automatically satisfied.

This construction also clarifies the compatibility between potentially unbounded projections and a bounded score head. Even when the manifold $\mathcal{M}$ or the projection values $P_{v_k}(z)$ are unbounded, choosing bounded slice functions such as the sigmoid ensures

$$0 \le g_{v_k}(t) \le 1, \qquad 0 \le h(z) \le 1,$$

for all $z \in \mathcal{M}$. Therefore, the linear-growth assumption on g_v is compatible with the probabilistic interpretation of h as a bounded score function.

C.5.2 Approximate Sliced Score Heads

The exact sliced representation in Theorem 4.5 can also be relaxed to an approximate one. Let $h_{\mathrm{mlp}} : \mathcal{M} \to [0,1]$ be a practical score head and suppose that there exists a sliced score head

$$\widetilde{h}(z) = \int_{S_o} g_v(P_v(z))\, d\lambda_o(v)$$

satisfying the assumptions of Theorem 4.5, such that on the union of group supports

$$\sup_{z \in \cup_s \mathrm{supp}(\mu_s)} \left| h_{\mathrm{mlp}}(z) - \widetilde{h}(z) \right| \le \eta.$$

Then, for any two groups $s, s' \in [A]$,

$$\begin{aligned}\left| \int h_{\mathrm{mlp}}\, d\mu_s - \int h_{\mathrm{mlp}}\, d\mu_{s'} \right| &\le \left| \int \widetilde{h}\, d\mu_s - \int \widetilde{h}\, d\mu_{s'} \right| + 2\eta \\ &\le L_{\mathrm{slice}} \left(\mathrm{CHSW}_p^p(\mu_s, \mu_{s'}) \right)^{1/p} + 2\eta.\end{aligned}$$

Consequently,

$$\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h_{\mathrm{mlp}} \circ f_\theta) \le 2L_{\mathrm{slice}} \pi_{\min}^{-1/p} U_{\mathrm{CHSW}}^{1/p} + 2\eta.$$

Thus, even when the practical MLP head is not exactly in sliced integral form, the CHSW-based soft-DP control remains informative whenever the MLP head is well approximated by a sliced score head on the learned representation support.

C.5.3 Relation to Standard MLP Score Heads

In our implementation, the classifier is a one-hidden-layer MLP that outputs two logits $(\ell_0(z), \ell_1(z))$. For binary classification, the class-one score is

$$h_{\mathrm{mlp}}(z) = \mathrm{softmax}(\ell_0(z), \ell_1(z))_1 = \sigma(\ell_1(z) - \ell_0(z)).$$

At evaluation time, with dropout disabled, the log-odds $\ell_1(z) - \ell_0(z)$ can be written as a finite superposition of ridge functions,

$$\ell_1(z) - \ell_0(z) = \sum_{k=1}^{K} c_k\, \rho(\langle w_k, z \rangle + b_k) + c_0.$$

Although the log-odds has an exact finite ridge-function form, the probability score h_{mlp} includes the outer sigmoid and therefore does not necessarily satisfy the additive sliced representation in Theorem 4.5 exactly. On compact or bounded representation supports, such MLP score heads can be approximated by sufficiently wide finite-slice score heads of the form (Barron, 1993; Pinkus, 1999; E et al., 2021):

$$\widetilde{h}(z) = \frac{1}{K} \sum_{k=1}^{K} g_{v_k}(P_{v_k}(z)),$$

where the directions v_k and scalar functions g_{v_k} play roles analogous to the weights w_k and activation-induced ridge functions in an MLP.

Therefore, the theorem applies directly to $\widetilde{h}$, and applies to the practical MLP head up to the approximation error

$$\eta = \sup_{z \in \cup_s \mathrm{supp}(\mu_s)} |h_{\mathrm{mlp}}(z) - \widetilde{h}(z)|.$$

More precisely, as proved in Appendix C.5.2, if h_{mlp} is uniformly approximated by such a sliced score head $\widetilde{h}$ on the learned representation support, then

$$\Delta_{\mathrm{DP}}^{\mathrm{soft}}(h_{\mathrm{mlp}} \circ f_\theta) \le 2L_{\mathrm{slice}} \pi_{\min}^{-1/p} U_{\mathrm{CHSW}}^{1/p} + 2\eta.$$

Thus, Theorem 4.5 should be interpreted as giving an exact guarantee for sliced score heads and an approximation-aware justification for standard MLP heads whose finite ridge-function structure is well aligned with the sliced representation class.

This connection also extends to multilayer MLP heads. A multilayer MLP can be viewed as a composition of finite ridge-type transformations, where each layer forms nonlinear combinations of one-dimensional projected features from the previous layer. Although a general multilayer MLP does not necessarily coincide exactly with the single-integral sliced form, it belongs to a closely related compositional extension of ridge-function models. Classical approximation results further show that sufficiently wide neural networks can represent broad classes of functions through such finite superpositions and their compositions (Barron, 1993; Pinkus, 1999; E et al., 2021).

Consequently, the finite-slice logistic head above is a direct instance of Theorem 4.5, while the MLP heads used in our experiments are approximation-compatible with the theorem through their ridge-function structure. This supports the practical feasibility of the sliced score-head assumption and explains why the CHSW-based bound remains relevant to common downstream score heads used in representation learning, without claiming that arbitrary MLP predictors exactly satisfy the theorem.

D Pseudocode and Computational Complexity

D.1 Pseudocode

We provide detailed pseudocode for our proposed method in Algorithm 1, which summarizes the complete end-to-end procedure and highlights the key computational steps and design choices of the algorithm. Since our experiments use the power $p = 2$, the pseudocode is presented for this setting.

D.2 Details about $\widetilde{U}_{\mathrm{CHSW}}$

D.2.1 What Algorithm 1 Actually Computes

The training objective minimized by Algorithm 1 is the following Monte Carlo approximation. For a minibatch with group-conditional empirical distributions $\{\hat{\mu}_s\}$, K random directions $\{v_k\}_{k=1}^K$ sampled from λ_o, and Q quantile levels $\{\tau_j\}_{j=1}^Q$, the algorithm accumulates

$$\widetilde{U}_{\mathrm{CHSW}} \;\approx\; \frac{1}{K}\sum_{k=1}^{K}\sum_{s}\pi_s\, W_p^p((P_{v_k})_{\#}\mu_s,\; \bar{\nu}_{v_k})\,,$$

where, for each direction v_k, $\bar{\nu}_{v_k}$ is the *one-dimensional W_p^p-barycenter* of $\{(P_{v_k})_{\#}\mu_s\}$ with weights $\{\pi_s\}$, whose quantile function satisfies

$$F^{-1}_{\bar{\nu}_v}(\tau) \;\in\; \arg\min_{q\in\mathbb{R}}\sum_{s}\pi_s\left|q - F^{-1}_{(P_v)_{\#}\mu_s}(\tau)\right|^p.$$

For $p = 2$, which is the setting used throughout our experiments, this minimizer admits the closed-form weighted average

$$F^{-1}_{\bar{\nu}_v}(\tau) \;=\; \sum_{s}\pi_s\, F^{-1}_{(P_v)_{\#}\mu_s}(\tau),$$

which is exactly the quantile barycenter computed in Step 17 of Algorithm 1. We therefore restrict the discussion in this section to $p = 2$.

In the population limit ($K \to \infty$, $Q \to \infty$), Algorithm 1 optimizes the *relaxed sliced barycentric objective*

$$\widetilde{U}_{\mathrm{CHSW}}(\mu_0,\dots,\mu_{A-1}) \;\triangleq\; \int_{S_o}\min_{\nu_v\in\mathcal{P}(\mathbb{R})}\sum_{s=0}^{A-1}\pi_s\, W_p^p((P_v)_{\#}\mu_s,\,\nu_v)\,d\lambda_o(v), \tag{29}$$

where for each direction v the minimizing $\nu_v = \bar{\nu}_v$ is chosen *independently*.

D.2.2 Distinction from U_{CHSW}

Recall that the CHSW barycenter regularizer defined in Section 4.1 is

$$U_{\mathrm{CHSW}}(\mu_0,\dots,\mu_{A-1}) \;\triangleq\; \min_{\nu\in\mathcal{P}_p(\mathcal{M})}\sum_{s=0}^{A-1}\pi_s\,\mathrm{CHSW}_p^p(\mu_s,\nu).$$

Algorithm 1 Training with CHSW Barycenter Regularization

Require: Training data $\{(x_i, y_i, s_i)\}_{i=1}^N$, encoder f_θ, classifier h_ψ, trade-off λ, #projections K, #quantiles Q, power p (default $p = 2$), manifold type $\mathcal{M}$

Initialize (θ, ψ)
for epoch $= 1, \dots, T$ **do**
  for minibatch $\mathcal{B}$ **do**
    Forward: $z_i \leftarrow f_\theta(x_i),\ \hat{y}_i \leftarrow h_\psi(z_i)$
    Prediction loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell(\hat{y}_i, y_i)$$

    Compute CHSW barycenter loss on the batch:
    Split by group $\mathcal{B}_s = \{i \in \mathcal{B} : s_i = s\}$ and let $\mathcal{S}_\mathcal{B} = \{s :\ |\mathcal{B}_s| > 0\}$
    Set barycenter weights as empirical group priors on the batch:

$$w_s \leftarrow \frac{|\mathcal{B}_s|}{|\mathcal{B}|}, \quad \forall s \in \mathcal{S}_\mathcal{B}$$

    Sample K random unit directions $\{v_k\}_{k=1}^K$
    Quantile levels $\tau_j = \frac{j}{Q+1},\ j = 1, \dots, Q$
    $\mathcal{L}_{\text{CHSW}} \leftarrow 0$
    for $k = 1, \dots, K$ **do**
      for each group s **do**
        1D manifold-aware projection:

$$u_{s,i}^{(k)} \leftarrow \Pi_\mathcal{M}(z_i; v_k), \quad i \in \mathcal{B}_s$$

        Empirical quantiles:

$$q_s^{(k)}(\tau_j) \leftarrow \text{Quantile}\big(\{u_{s,i}^{(k)}\}_{i \in \mathcal{B}_s}, \tau_j\big),\ \forall j$$

      end for
      Quantile barycenter:

$$\bar{q}^{(k)}(\tau_j) = \sum_s w_s\, q_s^{(k)}(\tau_j), \quad \forall j$$

      Accumulate deviation:

$$\mathcal{L}_{\text{CHSW}} \leftarrow \mathcal{L}_{\text{CHSW}} + \sum_s w_s \sum_{j=1}^{Q} \Big| q_s^{(k)}(\tau_j) - \bar{q}^{(k)}(\tau_j) \Big|^p$$

    end for
    Normalize:

$$\mathcal{L}_{\text{CHSW}} \leftarrow \frac{1}{KQ} \mathcal{L}_{\text{CHSW}}$$

    Joint objective and update:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda\, \mathcal{L}_{\text{CHSW}}$$

    Update (θ, ψ) by backpropagation on $\mathcal{L}$
  end for
end for
return trained (θ, ψ)

Let $\bar{\mu} \in \arg\min_{\nu \in \mathcal{P}_p(\mathcal{M})} \sum_s \pi_s\, \text{CHSW}_p^p(\mu_s, \nu)$ be any CHSW barycenter. Then

$$U_{\text{CHSW}} \;=\; \sum_{s=0}^{A-1} \pi_s \int_{S_o} W_p^p((P_v)_\# \mu_s,\ (P_v)_\# \bar{\mu})\, d\lambda_o(v).$$

The key structural difference is that in U_{CHSW} the reference measure projected onto each direction, namely $(P_v)_{\#}\bar{\mu}$, must arise as the pushforward of a *single manifold distribution* $\bar{\mu} \in \mathcal{P}_p(\mathcal{M})$. In $\widetilde{U}_{\mathrm{CHSW}}$, the reference measure $\bar{\nu}_v$ is chosen independently for each direction v and in general does not correspond to the pushforward of any single distribution on $\mathcal{M}$. Since $\widetilde{U}_{\mathrm{CHSW}}$ optimizes over a strictly larger feasible set, we immediately obtain the following inequality.

Proposition D.1 ($\widetilde{U}_{\mathrm{CHSW}} \le U_{\mathrm{CHSW}}$)**.** *For any $\mu_0, \dots, \mu_{A-1} \in \mathcal{P}_p(\mathcal{M})$,*

$$\widetilde{U}_{\mathrm{CHSW}}(\mu_0, \dots, \mu_{A-1}) \;\le\; U_{\mathrm{CHSW}}(\mu_0, \dots, \mu_{A-1}).$$

Proof. For the CHSW barycenter $\bar{\mu}$, choosing $\nu_v = (P_v)_{\#}\bar{\mu}$ for each v is a feasible (but generally suboptimal) choice in the direction-wise minimization defining $\widetilde{U}_{\mathrm{CHSW}}$. Therefore $\widetilde{U}_{\mathrm{CHSW}} \le U_{\mathrm{CHSW}}$. □

D.2.3 Fairness Guarantees Transfer to $\widetilde{U}_{\mathrm{CHSW}}$

Despite being a strict relaxation of U_{CHSW}, the objective $\widetilde{U}_{\mathrm{CHSW}}$ suffices to establish pairwise CHSW bounds, and hence all downstream fairness guarantees in Section 4.2 and Section 4.3, with U_{CHSW} replaced by $\widetilde{U}_{\mathrm{CHSW}}$ throughout.

Define, for each group s,

$$D_s \;\triangleq\; \left(\int_{S_o} W_p^p((P_v)_{\#}\mu_s,\, \bar{\nu}_v)\, d\lambda_o(v) \right)^{1/p}.$$

Since $\widetilde{U}_{\mathrm{CHSW}} = \sum_s \pi_s D_s^p$, we have $D_s^p \le \widetilde{U}_{\mathrm{CHSW}}/\pi_s$ for every s.

Lemma D.2 (Pairwise CHSW bound via $\widetilde{U}_{\mathrm{CHSW}}$)**.** *For all $s, t \in [A]$,*

$$\mathrm{CHSW}_p^p(\mu_s, \mu_t) \;\le\; \frac{2^p}{\pi_{\min}}\, \widetilde{U}_{\mathrm{CHSW}},$$

where $\pi_{\min} = \min_s \pi_s > 0$.

Proof. Fix any direction $v \in S_o$. By the triangle inequality for W_p on $\mathbb{R}$,

$$W_p((P_v)_{\#}\mu_s,\, (P_v)_{\#}\mu_t) \;\le\; W_p((P_v)_{\#}\mu_s,\, \bar{\nu}_v) + W_p((P_v)_{\#}\mu_t,\, \bar{\nu}_v)\,.$$

Integrating over $v \sim \lambda_o$ and applying the Minkowski inequality for $L^p(\lambda_o)$ on the right-hand side,

$$\mathrm{CHSW}_p(\mu_s, \mu_t) \;\le\; D_s + D_t.$$

Taking p-th powers and applying $(a+b)^p \le 2^{p-1}(a^p + b^p)$ for $a, b \ge 0$,

$$\begin{aligned} \mathrm{CHSW}_p^p(\mu_s, \mu_t) \;&\le\; 2^{p-1}(D_s^p + D_t^p) \\ &\le\; 2^{p-1}\left(\frac{1}{\pi_s} + \frac{1}{\pi_t}\right)\widetilde{U}_{\mathrm{CHSW}} \\ &\le\; \frac{2^p}{\pi_{\min}}\, \widetilde{U}_{\mathrm{CHSW}}. \end{aligned}$$

□

Lemma D.2 is structurally identical to Lemma 4.1, with $\widetilde{U}_{\mathrm{CHSW}}$ in place of U_{CHSW}. Since Theorem 4.4 and Theorem 4.5 depend on the pairwise CHSW bound only through Lemma 4.1, both theorems hold verbatim with U_{CHSW} replaced by $\widetilde{U}_{\mathrm{CHSW}}$.

Remark D.3. Proposition D.1 shows $\widetilde{U}_{\mathrm{CHSW}} \le U_{\mathrm{CHSW}}$, so minimizing $\widetilde{U}_{\mathrm{CHSW}}$ does not in general drive U_{CHSW} to zero. However, Lemma D.2 establishes that $\widetilde{U}_{\mathrm{CHSW}} \to 0$ is itself sufficient to guarantee $\mathrm{CHSW}_p^p(\mu_s, \mu_t) \to 0$ for all pairs s, t, which is the condition required by Theorem 4.4 and Theorem 4.5. Therefore, replacing U_{CHSW} with $\widetilde{U}_{\mathrm{CHSW}}$ does not weaken the downstream fairness guarantees; instead, it provides a tractable relaxed objective with direct and tighter pairwise-discrepancy control.

D.3 Computational Complexity

Let μ and ν be represented by empirical samples $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ on the manifold. Algorithm 1 approximates CHSW_p by Monte-Carlo sampling K slicing directions $v_k \sim \lambda_o$ and computing K one-dimensional Wasserstein distances between the projected scalars $P_{v_k}(x_i)$ and $P_{v_k}(y_j)$. For a fixed direction v, forming the projections costs $O!\big((n+m)C_{\mathrm{proj}}\big)$ where C_{proj} is the cost of evaluating the manifold-aware projection $P_v(\cdot)$ once, and the subsequent 1D OT computation is dominated by sorting (or equivalently quantile extraction) with cost $O(n\log n + m\log m)$. If a fixed grid of Q quantiles is used after sorting, the additional matching cost is $O(Q)$ and does not change the leading term. Therefore, the overall complexity scales as

$$T_{\mathrm{CHSW}} = O\Big(K\big((n+m)\,C_{\mathrm{proj}} + n\log n + m\log m\big)\Big),$$

with $O(n+m)$ memory if directions are processed in a streaming fashion (or $O(K(n+m))$ if all projections are stored). In our settings, C_{proj} is cheap for hyperbolic and pullback-Euclidean cases (closed-form scalar coordinate computations and elementwise chart maps followed by inner products, yielding $C_{\mathrm{proj}} = O(d)$ for embedding dimension d), while for the SPD-AIRM case the dominant cost comes from matrix factorizations (e.g., Cholesky/LDL-type decomposition), giving $C_{\mathrm{proj}} = O(r^3)$ for $r \times r$ SPD matrices. Importantly, this factorization is direction-independent and can be reused across the K directions, resulting in an effective projection-stage cost $O((n+m)r^3 + K(n+m)r)$. Overall, Algorithm 1 reduces manifold OT to K inexpensive 1D OT problems plus per-sample manifold projections, making CHSW computationally tractable and scalable linearly in the number of projections K (up to the sorting term).

E Experiments Details

In this section, we provide supplementary experimental details.

E.1 Consistent Experimental Protocol

To ensure a fair comparison, all methods are evaluated under a consistent experimental protocol whenever applicable. Specifically, we use the same data splits, preprocessing procedures, model backbones, evaluation metrics, and random seeds across different methods. For example, the LFWA dataset is split into training/validation/test sets with a ratio of $64/16/20$, and all trainable methods are optimized using Adam with a learning rate of 10^{-3} and a fixed training budget, typically 120 epochs across datasets.

For each method, we tune only the method-specific regularization strength or loss coefficient on the validation set and report the best-performing result under its own hyperparameter selection, following common practice in fair representation learning. This allows each method to be compared under its appropriate operating point while keeping the overall experimental setting consistent.

Some baselines, such as Wasserstein barycenter post-processing methods, follow a different training paradigm and therefore cannot be fully aligned with end-to-end trainable methods in terms of optimization procedure. Nevertheless, we keep the data splits, representation backbone, evaluation metrics, and reporting protocol consistent across all methods to ensure a fair and transparent comparison. More details are presented in the following sections.

E.2 Simulation Data Generation

E.2.1 Pullback Euclidean Space

We conduct Pullback Euclidean Space simulation experiments in a controlled binary classification setting with group-dependent distribution shift. A binary sensitive attribute $S \in \{0,1\}$ is first sampled, and the binary label $Y \in \{0,1\}$ is then generated from an S-dependent probabilistic mechanism, introducing a tunable group-level base-rate bias. To synthesize features containing both task-relevant and spurious components, we draw two random unit directions in $\mathbb{R}^d$: one aligned with the predictive signal associated with Y and the other aligned with the nuisance variation associated with S. In a latent Euclidean space, we construct a linear Gaussian model by combining these

directional components and adding isotropic Gaussian noise. The latent samples are then mapped to the observation space via a smooth coordinate-wise nonlinear transformation, so that distances in the observation space are measured through the pullback Euclidean metric induced by ϕ, i.e., $d(x, x') = ||\phi(x) - \phi(x')||$ with $\phi(\cdot) = \text{asinh}(\cdot)$.

E.2.2 HYPERBOLIC SPACE

In the Hyperbolic Space simulation experiments, we generate a controlled binary classification dataset directly on the Poincaré ball model with curvature -1, thereby inducing a native hyperbolic geometric structure. We first sample a binary sensitive attribute $S \in \{0, 1\}$ and then generate the binary label $Y \in \{0, 1\}$ from an S-dependent probabilistic mechanism, which introduces a tunable group-level base-rate disparity (bias strength $= 2.2$). To create a clear geometric separation between groups, we specify two group centers in hyperbolic space: the center of group $S = 0$ is placed at the origin, while the center of group $S = 1$ is positioned at a prescribed hyperbolic distance from the origin (inter-center distance $= 3.0$). For each sample, we construct a tangent-space perturbation at the corresponding group center by combining an isotropic Gaussian component with a label-dependent signal component, and map it back to the Poincaré ball via the exponential map to obtain the observed feature $X \in \mathbb{R}^d$ (we use $n = 10{,}000$, $d = 12$; within-group noise scale $= 0.5$; signal strength $= 2.3$).

E.2.3 SPD AIRM SPACE

In the SPD–AIRM simulation experiments, we generate a controlled binary classification dataset directly on the manifold of symmetric positive definite matrices equipped with the affine-invariant Riemannian metric, thereby inducing an intrinsic non-Euclidean geometric structure. We first sample a binary sensitive attribute $S \in \{0, 1\}$ with $S \sim \text{Bernoulli}(0.5)$, and then draw the binary label $Y \in \{0, 1\}$ from an S-dependent logistic mechanism to introduce a tunable group-level base-rate disparity (group bias strength $= 2.2$). To create sensitive-attribute-driven spurious geometric separation, we specify two group centers on the SPD manifold: the center for $S = 0$ is set to $C_0 = I_m$, while the center for $S = 1$ is constructed as $C_1 = \exp(\text{spurious_strength} \cdot D)$ where D is a randomly sampled symmetric direction matrix (spurious strength $= 3.0$). Each sample uses C_S as its base point. For each instance, we form a symmetric tangent-space perturbation by combining isotropic Gaussian noise with a label-dependent signal aligned with class-specific symmetric directions (noise scale $= 0.8$; signal strength $= 2.3$), and map it back to the manifold via the AIRM exponential map to obtain an SPD observation $X \in \mathbb{S}^m_{++}$. Finally, we vectorize each SPD matrix by extracting its lower-triangular entries to produce the Euclidean feature vector used for learning. We use $n = 1000$ samples with SPD size $m = 3$ (yielding $d = m(m + 1)/2 = 6$).

E.3 IMPLEMENTATION DETAILS

E.3.1 MANIFOLD SELECTION

In practice, the ground-truth geometry of the learned representation space is usually unknown. Thus, the choice of manifold is often treated as an empirical modeling decision rather than a directly observable property of the data. Prior geometry-aware representation learning follows a similar protocol: evaluating representative candidate geometries and selecting the one that performs best for the given task and data structure (Bachmann et al., 2020; Chami et al., 2019; Gu et al., 2018).

Following this practice, we evaluate our method on several representative geometries, including Euclidean/pullback Euclidean, hyperbolic, and SPD manifolds. The empirical results show that non-Euclidean geometries can improve over purely Euclidean representations in settings where the underlying data structure is better captured by curved spaces. This observation supports the practical relevance of incorporating manifold geometry into fair representation learning.

Importantly, our goal is not to tune the manifold geometry using fairness metrics. Rather, we use these representative geometries to examine whether the proposed CHSW-based regularization remains effective across different geometric settings. The results demonstrate that our method is not tied to a single manually selected manifold, but can be applied consistently across several commonly used non-Euclidean representation spaces.

E.3.2 NEURAL ARCHITECTURES (ENCODER f_θ AND PREDICTOR h_ψ)

In all experiments, we use the same backbone network for the representation encoder and predictor to ensure comparable model capacity across methods. Only the training objective (i.e., the fairness regularizer) differs. Concretely, the encoder f_θ is implemented as a 3-layer MLP, producing a latent embedding $z \in \mathbb{R}^{d_z}$. The predictor h_ψ is a 2-layer MLP, trained with cross-entropy on logits for classification.

E.3.3 CHOICE OF ORIGIN FOR CHSW

For manifolds without global translation invariance, the CHSW distance in Definition 3.4 is defined relative to a fixed base point (origin), which determines both the tangent space from which slicing directions are sampled and the reference point for geodesic or horospherical projections. In our implementation, this origin is fixed a priori and kept constant throughout training and evaluation, ensuring determinism and reproducibility.

Specifically, for each manifold considered in our experiments, we adopt a canonical choice of origin:

(1) Hyperbolic Space (Poincaré Ball Model). We fix the origin at the ball center 0. Latent representations are interpreted as elements of the tangent space $T_0\mathbb{H}^d$ and are mapped to the manifold via the exponential map $\exp_0$. Slicing directions are sampled as unit vectors in $T_0\mathbb{H}^d$, and all geodesic or horospherical projections are defined with respect to geodesics or Busemann functions anchored at the origin. The origin is not updated or re-centered during training.

(2) SPD Manifold with AIRM Geometry. We fix the base point at the identity matrix I. Latent vectors are mapped to symmetric matrices in the tangent space $T_I\mathcal{S}^m_{++}$ and then transported to the manifold via the matrix exponential. Slicing directions are sampled in the corresponding tangent parameterization, and projections are computed using horospherical coordinates derived relative to the identity. As in the hyperbolic case, the origin remains fixed throughout training.

(3) Pullback Euclidean Space. For pullback manifolds defined via a global diffeomorphism ϕ, all projections are performed in the Euclidean space induced by ϕ. In this setting, no explicit origin selection is required beyond the fixed coordinate chart.

By fixing a canonical origin for each manifold and sampling all slicing directions in the corresponding tangent space, our implementation avoids ambiguity in the definition of CHSW and guarantees full reproducibility. We empirically examine alternative choices such as random and data-adaptive base points in Appendix E.5, and find that simple canonical origins provide a stable and reproducible default in practice.

E.3.4 MANIFOLD-AWARE PROJECTIONS IN CHSW

The Cartan–Hadamard Sliced–Wasserstein (CHSW) distance (Definition 3.4) is defined via one-dimensional manifold-aware projections P^v along directions v sampled from the unit sphere of the tangent space at a chosen origin. While both geodesic and horospherical (Busemann) projections are admissible in theory, the specific variant depends on the underlying manifold and its geometric structure. Below we detail the exact projection schemes used in our experiments.

(1) Hyperbolic Space. For hyperbolic embeddings, we adopt the geodesic projection coordinate, corresponding to GCHSW. We fix the origin at the center of the Poincaré ball $o = 0$, and sample slicing directions v from the unit sphere of the tangent space $T_o\mathbb{H}^d$. Given a point $x \in \mathbb{H}^d$, the projection $P^v(x)$ is defined as the signed coordinate of x along the geodesic

$$\gamma_v(t) = \exp_o(tv),$$

i.e., the scalar parameter t such that the orthogonal projection of x onto γ_v is $\gamma_v(t)$. This geodesic coordinate admits a closed-form expression in hyperbolic space and is computed in our implementation via equivalent formulas derived from the Poincaré/Lorentz models. The origin remains fixed throughout training and evaluation.

Dataset	Origin	Description	ACC ↑	$\Delta_{DP}\downarrow$	$\Delta_{DP}^{\text{soft}}\downarrow$
Hyperbolic	center ($\|o\| = 0$)	canonical origin (Poincaré ball center)	**0.521±0.061**	0.245±0.158	**0.005±0.006**
	$\|o\| = 0.15$	random point with small displacement	0.502±0.065	0.236±0.153	0.009±0.008
	$\|o\| = 0.30$	random point with larger displacement	0.477±0.052	**0.226±0.160**	0.011±0.008
	data_mean	heuristic: Fréchet mean of all samples	0.474±0.052	0.245±0.142	0.010±0.008
	group_barycenter	adaptive: fairness-aware group barycenter	0.504±0.064	0.234±0.152	0.009±0.008
SPD	identity (I)	canonical origin (identity matrix)	**0.795±0.034**	0.210±0.145	0.079±0.068
	random1	random SPD matrix (seed=42)	0.778±0.044	0.188±0.197	0.081±0.087
	random2	random SPD matrix (seed=123)	0.793±0.037	**0.179±0.159**	**0.073±0.076**
	data_mean	heuristic: Fréchet mean of all samples	0.789±0.050	0.249±0.167	0.120±0.077
	group_mean	adaptive: group-wise barycenter	0.790±0.047	0.263±0.169	0.125±0.082

Table 3: Sensitivity to the Choice of Origin o. The best results are highlighted in bold.

(2) SPD Manifold with Affine-Invariant Metric. For symmetric positive definite (SPD) matrices equipped with the affine-invariant Riemannian metric (AIRM), we use a horospherical, or Busemann-type, projection formulation of CHSW, yielding an HCHSW-style discrepancy. The origin is fixed at the identity matrix $o = I$. Slicing directions are taken in the tangent space $T_I\mathcal{S}_{++}^m$, which can be identified with the space of symmetric matrices endowed with the Frobenius inner product.

For a unit direction $V \in T_I\mathcal{S}_{++}^m$, let

$$\gamma_V(t) = \mathrm{Exp}_I(tV) = \exp(tV), \qquad t \geq 0,$$

be the corresponding AIRM geodesic ray. The ideal horospherical projection is defined by the Busemann function associated with this ray:

$$P_{\mathrm{hor}}^V(X) = b_V(X) := \lim_{t\to\infty} \left[d_{\mathrm{AIRM}}\big(X, \gamma_V(t)\big) - t\right],$$

where

$$d_{\mathrm{AIRM}}(X,Y) = \left\|\log\left(X^{-1/2}YX^{-1/2}\right)\right\|_F.$$

(3) Pullback Euclidean Manifold. For pullback manifolds defined by a global diffeomorphism $\phi : \mathcal{M} \to \mathbb{R}^d$, we perform all slicing operations in the induced Euclidean space. In this setting, the pullback metric satisfies

$$g_x(u,v) = \langle \phi_{*,x}u, \phi_{*,x}v\rangle, \qquad d_{\mathcal{M}}(x,y) = \|\phi(x) - \phi(y)\|,$$

implying that geodesics in $\mathcal{M}$ correspond to straight lines in the ϕ-coordinates. Consequently, both geodesic and horospherical projections reduce to standard linear projections, and the CHSW distance degenerates to the classical sliced-Wasserstein distance applied to the pushforward measures (Bonet et al., 2025, Lemma 25):

$$\mathrm{CHSW}_p(\mu,\nu) = \mathrm{SW}_p(\phi_{\#}\mu, \phi_{\#}\nu).$$

In practice, we compute $P^v(x) = \langle v, \phi(x)\rangle$ with directions v sampled uniformly from the Euclidean unit sphere. No additional origin-dependent construction is required in this case.

E.3.5 Computing Infrastructure.

All experiments were conducted on a server equipped with NVIDIA TITAN V GPUs and Intel(R) Xeon(R) E5-2620 v4 CPUs at 2.10 GHz, with 173 GiB of system memory. Each experimental run used a *single* NVIDIA TITAN V GPU with 12 GB of GPU memory. The server ran Ubuntu 16.04.6 LTS and NVIDIA driver 460.80, with `nvidia-smi` reporting CUDA 11.2 support. Our implementation used Python 3.8.5 and PyTorch 1.7.0 (`1.7.0+cu101`), built with CUDA 10.1. All compared methods were evaluated under the same hardware and software environment.

E.3.6 Hyperparameter Settings

For our method, we use CHSW projection directions $K = 120$ for all datasets. The origin o is chosen as the canonical origin of each underlying manifold. Specifically, for hyperbolic representations, we use the center of the Poincaré ball, i.e., $\|o\| = 0$; for SPD representations, we use the identity matrix I as the origin.

Dataset	K	ACC ↑	Δ_{DP} ↓	$\Delta_{DP}^{\text{soft}}$ ↓
Pullback	60	**0.819±0.019**	0.293±0.051	0.257±0.057
	120	0.758±0.064	**0.187±0.122**	**0.176±0.096**
	240	0.807±0.040	0.268±0.088	0.243±0.075
Hyperbolic	60	0.485±0.055	**0.229±0.159**	0.013±0.009
	120	**0.521±0.061**	0.245±0.158	**0.005±0.006**
	240	0.513±0.073	0.281±0.160	0.008±0.006
SPD	60	0.785±0.048	0.221±0.130	0.088±0.057
	120	**0.795±0.034**	**0.210±0.145**	0.079±0.068
	240	0.793±0.034	0.221±0.137	**0.078±0.066**
Adult	60	**0.843±0.005**	0.115±0.008	0.098±0.017
	120	0.840±0.007	0.098±0.034	0.082±0.033
	240	0.823±0.040	**0.081±0.054**	**0.067±0.048**
LFWA	60	0.902±0.011	0.059±0.010	0.058±0.009
	120	**0.905±0.004**	**0.048±0.012**	**0.046±0.010**
	240	0.902±0.005	0.057±0.009	0.051±0.009

Table 4: Sensitivity to the Number of CHSW Projection Directions K. The best results are highlighted in bold.

The representation dimension d_z and hidden dimension h are set as follows: $d_z = 6$ and $h = 64$ for Pullback, $d_z = 8$ and $h = 64$ for Hyperbolic, $d_z = 6$ and $h = 64$ for SPD, $d_z = 16$ and $h = 64$ for Adult, and $d_z = 16$ and $h = 64$ for LFWA. The learning rate is fixed to 10^{-3} across all datasets.

For all baseline methods, we follow the hyperparameter settings recommended in their original papers.

E.4 Representation and Prediction Distributions Analysis

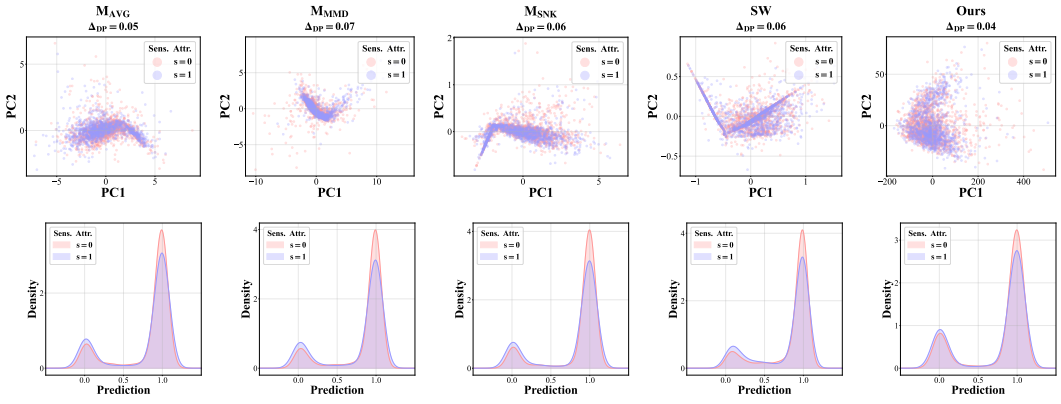


Figure 5: Representation (top row) and prediction distributions (bottom row) analysis on the LFWA dataset. The sensitive attribute (s) indicates whether "WavyHair" is present in the image (s=0) or absent (s=1).

Figure 5 presents a qualitative analysis on the LFWA dataset, including PCA visualizations of the learned representations (top row; scatter plots of PC1 vs. PC2) and the corresponding prediction distributions (bottom row; kernel density estimates). The sensitive attribute s indicates whether WavyHair is present in the image ($s = 0$) or absent ($s = 1$). As can be seen from the figure, our method yields substantially more overlapping distributions between the two demographic groups, both in the representation space and in the prediction space. This improved alignment suggests reduced sensitivity-dependent discrepancy, leading to better fairness and a smaller Δ_{DP}.

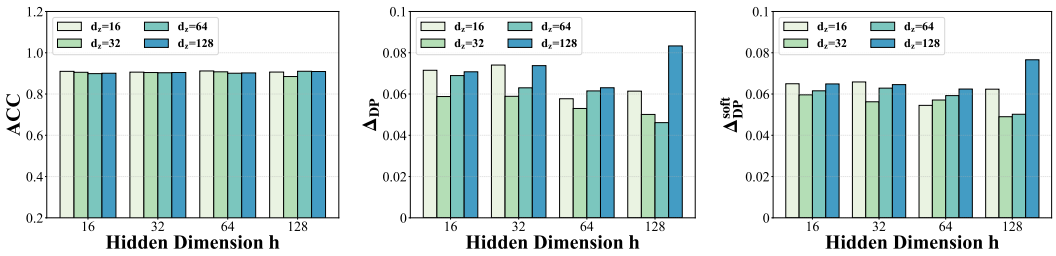


Figure 6: Sensitivity to the representation dimension d_z on the LFWA dataset.

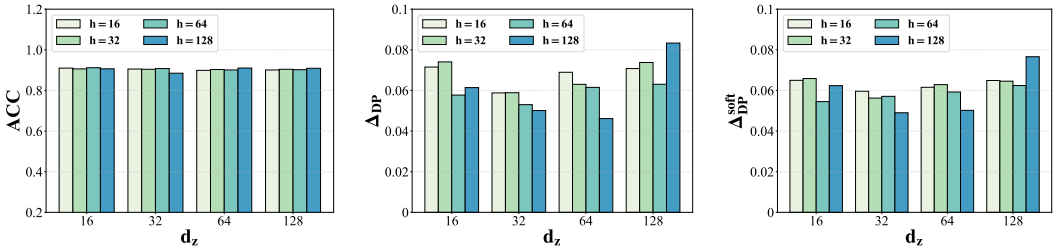


Figure 7: Sensitivity to the hidden dimension h on the LFWA dataset.

E.5 Hyperparameter Sensitivity Analysis

Sensitivity to the representation dimension d_z and the hidden dimension h. We conduct a comprehensive hyperparameter analysis of our proposed method on the LFWA dataset, systematically examining the effects of two key hyperparameters: the representation dimension d_z (ranging from 16 to 128) and the hidden dimension h (varying from 16 to 128) on model performance. As illustrated in Figure 6 and Figure 7 reports the hyperparameter sensitivity analysis on the LFWA dataset, where we vary the representation dimension d_z from 16 to 128 and the hidden dimension h from 16 to 128. Overall, the performance trends remain relatively stable across a broad range of settings, suggesting that our method is robust to reasonable hyperparameter choices. Meanwhile, we observe that extremely small or extremely large values of h and d_z can lead to increased Δ_{DP} and $\Delta_{\mathrm{DP}}^{\mathrm{soft}}$. This is likely because overly small dimensions limit the model's representational capacity, whereas overly large dimensions can exacerbate overfitting and consequently amplify group-dependent discrepancies.

Sensitivity to the Number of CHSW Projection Directions K. We further evaluate the sensitivity to the number of projection directions K, as summarized in Table 4. Overall, the results are stable across different values of K, with moderate values (e.g., 120) achieving a good balance between performance and efficiency. This demonstrates that our method is not overly sensitive to these hyperparameters and can be applied with simple default choices.

Sensitivity to origin choice o in CHSW. Our implementation follows the standard CHSW formulation with a fixed reference origin (e.g., ball center or identity)(Bonet et al., 2025). We further evaluate multiple alternatives, including random and data-adaptive choices. As shown in Table 3, although different origins lead to some variation, the overall performance remains stable, and our method consistently improves fairness metrics. Therefore, CHSW is robust to the choice of origin in practice, and simple canonical choices already work well.

E.6 Comparison with adversarial FRL baselines

Adversarial fair representation learning (FRL) methods and our CHSW-based approach are based on fundamentally different paradigms. Adversarial methods, such as LAFTR (Madras et al., 2018), enforce fairness through a minimax game between a predictor and an adversary that attempts to recover the sensitive attribute from the learned representation. In contrast, our method directly

Dataset	Metric	CHSW	LAFTR
Adult	ACC ↑	**0.840±0.007**	0.835±0.003
	Δ_{DP} ↓	**0.098±0.034**	0.168±0.009
	$\Delta_{DP}^{\text{soft}}$ ↓	**0.082±0.033**	0.176±0.003
LFWA	ACC ↑	**0.905±0.004**	0.893±0.004
	Δ_{DP} ↓	**0.048±0.012**	0.074±0.006
	$\Delta_{DP}^{\text{soft}}$ ↓	**0.046±0.010**	0.068±0.004

Table 5: Comparison with an Adversarial FRL Baseline on Adult and LFWA. The best results are highlighted in bold.

imposes distributional alignment between group-conditional representation distributions through a geometry-aware discrepancy defined on the underlying manifold.

These differences lead to distinct optimization objectives, inductive biases, and training dynamics. Adversarial methods often require careful balancing between the prediction loss and the adversarial objective, whereas our formulation provides a direct discrepancy-minimization objective through CHSW. To contextualize the empirical performance of our method, we additionally compare against LAFTR as a representative adversarial FRL baseline on Adult and LFWA. As shown in Table 5, CHSW achieves better fairness on both Δ_{DP} and $\Delta_{DP}^{\text{soft}}$ while maintaining competitive or better accuracy, further supporting the effectiveness of CHSW for fairness-aware representation learning.

F LIMITATIONS

Although our framework relaxes Euclidean assumptions by extending fair representation learning and distributional discrepancy measures to Hadamard manifolds, some theoretical questions remain open. On general Hadamard manifolds, CHSW is a pseudodistance rather than a full metric, so our current guarantees apply to the soft demographic-parity violation $\Delta_{DP}^{\text{soft}}$ instead of the hard-threshold violation Δ_{DP}. Empirically, reducing $\Delta_{DP}^{\text{soft}}$ often also improves Δ_{DP}, suggesting that the soft criterion is practically meaningful, while a strict hard-DP guarantee remains an interesting direction for future work. In addition, our analysis relies on the nonpositive-curvature structure of Hadamard manifolds; extending the framework to positive-curvature, mixed-curvature, product, or non-smooth spaces would further broaden its applicability.