Fair Data Representation for Machine Learning at the Pareto Frontier

Shizhou Xu SHZXU@UCDAVIS.EDU
Department of Mathematics
University of California Davis
Davis, CA 95616-5270, USA

Thomas Strohmer STROHMER@MATH.UCDAVIS.EDU
Department of Mathematics
Center of Data Science and Artificial Intelligence Research
University of California Davis
Davis, CA 95616-5270, USA

Abstract

As machine learning powered decision-making becomes increasingly important in our daily lives, it is imperative to strive for fairness in the underlying data processing. We propose a pre-processing algorithm for fair data representation via which supervised learning results in estimations of the Pareto frontier between prediction error and statistical disparity. In particular, the present work applies the optimal affine transport to approach the post-processing Wasserstein barycenter characterization of the optimal fair L^2-objective supervised learning via a pre-processing data deformation. Furthermore, we show that the Wasserstein geodesics from the conditional (on sensitive information) distributions of the learning outcome to their barycenter characterize the Pareto frontier between L^2-loss and the average pairwise Wasserstein distance among sensitive groups on the learning outcome. Numerical simulations underscore the advantages: (1) the pre-processing step is compositive with arbitrary conditional expectation estimation supervised learning methods and unseen data; (2) the fair representation protects the sensitive information by limiting the inference capability of the remaining data with respect to the sensitive data; (3) the optimal affine maps are computationally efficient even for high-dimensional data.

Keywords: statistical parity, equalized odds, Wasserstein barycenter, Wasserstein geodesics, conditional expectation estimation

1. Introduction

Our society is increasingly influenced by artificial intelligence as (direct or indirect) decision-making processes become more reliant on statistical inference and machine learning. The potentially significant long-term impact from sequences of automated (facilitate of) decision-making has brought large concerns about bias and discrimination in machine learning [5, 38]. Machine learning based on unbiased algorithms can naturally inherit the historical biases that exist in data and hence reinforce the bias via automated decision-making process [12].

One straightforward partial remedy is to exclude the sensitive variables from the data set used in the learning and decision process. But such exclusion merely eliminates disparate treatment, which refers to direct discrimination, and leaves disparate impact, which refers to unintended or indirect discrimination, remaining in both data and learning outcome [23].

Examples of the legal doctrine of disparate impact include Griggs v. Duke Powers Co. [10] and Ricci v. DeStefano [1], where the decision is based on factors that are strongly correlated with race, such as intelligence qualification in the former and the racially disproportionate test result in the latter, are ruled illegal by the US supreme court. As a result, along with the trending development of automated decision making, the need for more sophisticated but practical techniques has made fairness in machine learning an important research area [33].

Two important but potentially conflicting goals of fair machine learning are *group fairness*, which aims to achieve similarity in predictions conditioned on sensitive information, and *individual fairness*, which aims for similar treatment of similar individuals regardless of the sensitive information. The present work targets an important definition in group fairness: *statistical parity* [21], because it is closely related to disparate impact and hence long-term structural influence [45], while individual fairness focuses more on the short-term individual consequence. In the remainder of this paper, fairness and statistical parity are used interchangeably[1].

Before further discussing statistical parity, we note that fairness in machine learning should not be defined by a single condition without considering the application context. The goal of the present work is to provide theoretically reliable and explainable tools to help practitioners obtain the optimal (w.r.t. utility) solutions at any chosen statistical disparity level, provided one chooses to adopt statistical parity (or limited statistical dependence between the learning outcome and the sensitive information) as a meaningful fairness definition in one's particular application context.

Remark 1.1 below provides a more detailed discussion on statistical parity, namely how the utility optimization solves some major insufficiency of the original statistical parity definition and improves statistical parity to proportional equality, a fairness concept similar to equity in modern ethics which can be traced back to Aristotle and Plato [6, 19].

Remark 1.1 (Statistical parity enhanced by utility optimization) *Statistical parity is one of the most important definitions of group fairness. It has advantages such as (1) legal support on mitigating adverse impact and (2) the long-term effect resulting from the enforced involvement of minority groups or diversity in learning outcome via affirmative action [27]. On the other hand, there are three major criticisms about statistical parity that are often mentioned, e.g. see [21, 26]: (1) reduced utility, (2) self-fulfilling prophecy, (3) subset targeting. However, we notice that the first two are insufficiencies with respect to utility. Therefore, the proposed method mitigates these two insufficiencies.*

1 *(Utility) The development of the Pareto frontier allows us to achieve a desirable statistical disparity level with theoretically provable minimum (hence necessary) utility sacrifice. Equivalently, practitioners can choose a tolerable utility sacrifice level so that the Pareto frontier will provide a learning outcome with the minimum statistical disparity while not violating the utility sacrifice tolerance.*

2 *(Self-fulfilling prophecy) As mentioned in [21, 26], self-fulfilling prophecy results from random, careless, or malicious selection in minority groups. But the barycenter char-*

[1] There are many other notions of fairness, such as equalized odds or equal opportunity, which all have their benefits and shortcomings [16]. A discussion of the advantages or disadvantages of the different concepts of fairness is beyond the scope of this paper.

> *acterization method guarantees the optimal fair model to make good selections in all sensitive groups to maximize utility. Section 1.2 contribution point 4 and Section 2.1 provides, respectively, the intuitive and technical explanation of how the utility maximization enforces the model to give similar learning outcomes to data points sharing relatively (within their sensitive groups) similar qualifications. For example, if race is the sensitive information and an admission test score is the only qualification variable, a barycenter-characterized optimal fair admission model would give admission to the same percentage of top-score students in each of their racial groups.*

Interestingly, the interpretation is consistent with the philosophical definition of fairness involving proportional equality: a model is fair (with respect to the sensitive information) if it distributes proportional chance or prediction to proportionally qualified independent variables within each of the sensitive groups.

Beginning with [21], there is now a sizable body of research studying fair machine learning solutions. The resulting approaches can be categorized into the following: (1) preprocessing: deform data before training to mitigate sensitive information in the learning outcome [13, 29]; (2) in-processing: implement the definition of fairness in the training process by penalizing unfair outcome [8, 43]; (3) post-processing: enforce the definition of fairness directly on the learning outcome [26, 28].

In recent years, the post-processing approach has received significant attention due to the following remarkable result: the optimal fair distribution of supervised learning, such as classification [28] and regression [18, 24], can be characterized as the Fréchet mean of the learning outcome marginals on the Wasserstein space, which is also known as the Wasserstein barycenter in the optimal transport literature. (See Remark 2.2 for more details on learning outcome marginals.) The following remark provides an intuition of the Wasserstein ($\mathcal{W}_2$) barycenter characterization, on which we develop our theoretical results and algorithms.

Remark 1.2 (Intuition of Wasserstein barycenter characterization) *The Fréchet mean is the closest point to a set of points in a metric space and, therefore, a generalization of the mean on the Euclidean space to general metric spaces such as the Wasserstein space. Intuitively, one can consider the barycenter (Fréchet mean in Wasserstein space) characterization of optimal fair learning outcome as an analog of representing a set of points by their average, which thereby optimally (with respect to total moving distance) removes the disparity among those points, except that each point is now in Wasserstein space, and hence a distribution. See Section 1.2 contribution point 4 below for more details.*

Despite the theoretical elegance of the post-processing barycenter characterization, challenges remain in theory and practice (see Section 1.2 for a detailed explanation of the challenges), especially compared to pre-processing or data representation methods.

Fair machine learning using a pre-processing approach has been considered in [13, 23, 25, 37, 29]. While the Wasserstein barycenter provides a mathematically rigorous characterization of the post-processing optimal learning outcome, optimal fair data representation for general supervised learning models still lacks a theoretical characterization. See, for example, [16, Section 3.4, 3.5] for more details on the current challenges in fair data representation design for general machine learning models beyond classification, not to mention data representations that provide the optimal trade-off between accuracy and fairness.

The goal of the present work is to develop an optimal fair data representation characterization so that any supervised learning model, which aims to estimate the conditional expectation, trained via the fair data representation results in a fair estimation of the post-processing Wasserstein barycenter characterized optimal fair learning outcome. The ultimate goal is to develop a method that enjoys both the mathematically rigorous characterization of post-processing and the flexibility of pre-processing.

1.1 Optimization Problems with Sensitive Variable Independence Constraint

The statistical parity constraint for supervised learning or data representation in a nutshell is a constraint on the dependence between the learning outcome and a chosen sensitive variable. Equivalently, the constraint limits the ability to access or reverse engineer the sensitive variable from the learning outcome or data representation. Therefore, although the theory and methods in the present work aim to solve current challenges in machine learning fairness, they can also be useful in other areas where sensitive or undesirable information needs to be eliminated within the existing learning outcome or data. One example of such an area other than fair machine learning is machine (feature) unlearning. It starts from [14] and now has a sizable body of research works.

Here, we summarize the constrained optimization problems solved in the present work. We prove existence (and uniqueness, if possible) results via a constructive characterization approach so that an explicit formula of the solutions becomes available. Practitioners and researchers interested in limiting the statistical dependence between the learning outcome or data representation and certain feature variables can directly refer to the corresponding section for results. We leave the underlying motivations resulting from machine learning fairness to the following two subsections.

In Section 3, we target the following problem:

Problem 1 (Optimal fair L^2-objective learning outcome)

$$\inf_{f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})} \{||Y - f(X,Z)||_2^2 : f(X,Z) \perp Z\} \tag{1}$$

Here, Y is the dependent variable, and $f(X,Z)$ is an estimator that uses the independent variable X and sensitive variable Z to estimate Y. The loss function aims to maximize utility by minimizing the L^2-norm between Y and $f(X,Z)$:

$$||Y - f(X,Z)||_2^2 = \int_\Omega ||Y - f(X,Z)||^2 d\mathbb{P}.$$

$(\Omega,\Sigma,\mathbb{P})$ is a probability space. For $S \in \{X,Y,Z\}$, $S : \Omega \to \mathcal{S}$ is a random variable (equivalently a measurable function) from Ω to the state space $\mathcal{S}$. $||\cdot||$ denotes the Euclidean norm. The constraint $f(X,Z) \perp Z$ guarantees that the final result is independent of the sensitive information Z and hence satisfies statistical parity. Finally, the admissible function space $L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})$ is the space of all square-integrable measurable functions from $\mathcal{X}\times\mathcal{Z}$ to $\mathcal{Y}$. (Our proof shows Problem 1 does not change if one allows all measurable functions

$\mathcal{X} \times \mathcal{Z}$ to $\mathcal{Y}$.) The reason of allowing all measurable functions in our problem setting is due to the recent development of deep neural networks that are capable of estimating arbitrary measurable functions.

In Section 4, we relax the above strict independence constraint by applying a quantification of statistical disparity: the *Wasserstein disparity*, which is the average pairwise Wasserstein distance among conditional (on Z) distributions of $f(X,Z)$, denoted by $D(f(X,Z),Z)$. It has the following desirable properties: (1) $D(f(X,Z)) = 0$ if and only if $f(X,Z) \perp Z$. (2) The larger D is, the more disparities there are among the marginals (w.r.t. Z) of $f(X,Z)$. (3) D has a meaningful interpretation in physics as the minimum expected amount of work required to remove the distributional discrepancy between two randomly chosen sensitive groups on the learning outcome. Therefore, fixing a disparity tolerance level $d \in [0,\infty)$,

Problem 2 (Optimal L^2-objective learning Pareto frontier)

$$\inf_{f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})} \{||Y - f(X,Z)||_2^2 : D(f(X,Z),Z) < d\} \tag{2}$$

gives us the corresponding Pareto optimal solution. That is, if one wants a lower L^2-loss than provided by the infimum in Problem 2, then it is necessary to increase the tolerance level d. Equivalently, if one wants to lower the tolerance level d, then it is necessary to sacrifice more L^2-loss than the infimum.

In Section 5, we provide a theoretical characterization of the solution to

Problem 3 (Optimal fair data representation for conditional expectation estimation)

$$\inf_{(\tilde{X},\tilde{Y}) \in \mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : \tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X},Z) \perp Z\}, \tag{3}$$

where $\mathcal{D}$ is the admissible data representation set we define later. Here, the objective function aims to maximize the potential utility remaining within the deformed data $(\tilde{X},\tilde{Y})$ by minimizing the L^2 distance between the perfect estimator $\mathbb{E}(\tilde{Y}|\tilde{X})$ on $(\tilde{X},\tilde{Y})$ and the original Y, so that better estimation of $\mathbb{E}(\tilde{Y}|\tilde{X})$ leads to better prediction of Y. The constraint $\tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X},Z) \perp Z$ guarantees: (1) $f(\tilde{X}) \perp Z$ for $\forall f : \mathcal{X} \to \mathcal{Y}$, such that any estimator of $E(\tilde{Y}|\tilde{X})$ is independent of Z; (2) The perfect adversarial estimator $\mathbb{E}(\tilde{Y}|\tilde{X},Z)$ is independent of Z, so that a better estimation of $E(\tilde{Y}|\tilde{X},Z)$ leads to more independence of Z (alignment between the training objective and independence constraint). In addition, one may choose the following alternative constraints according to the application context: (1) $\tilde{X} \perp Z$, which guarantees $f(\tilde{X}) \perp Z$ for all measurable f as mentioned above; (2) $(\tilde{X},\tilde{Y}) \perp Z$, which guarantees any (adversarial) supervised or unsupervised learning on $(\tilde{X},\tilde{Y})$ to be independent of Z. The first alternative is useful if only measurable functions of X are allowed, whereas the second should be applied when one does not know which features are dependent or independent. See Section 1.3 for a more detailed derivation and explanation of the data representation objective function and constraints.

1.2 Challenges and Contributions in Machine Learning Fairness

Now, we go back to the motivation behind the optimization problems listed above: fair machine learning. We first summarize the limitations of the current post-processing characterization and the current methods based on it to estimate the optimal fair learning outcome.

1. The post-processing barycenter characterization lacks theoretical and computational generalization to high-dimensional data spaces, such as text or image spaces. From a theoretical perspective, the current works [18, 24, 37] focus on classification and 1-dimensional regression. From a computational perspective, the current works apply the coupling of cumulative distribution functions (cdf) of the learning outcome sensitive conditionals to find the barycenter and the inverse of the cdf to compute the optimal transport map. Both the coupling and the inverse of the cdf are computationally expensive. Furthermore, since the inverse of the cdf cannot be generalized to high-dimensional spaces, the current methods lack the generalization to supervised learning with high-dimensional dependent variables.

   Due to the recent development of generative AI models, it is now important to have fair machine learning methods for arbitrarily high-dimensional data. We hope the present work on the L^2 space can be a starting point for fair machine learning or data representation on more general spaces for high-dimensional data.

2. The current post-processing barycenter characterization lacks both theoretical and computational generalization to (an estimation of) the optimal trade-off, also known as the Pareto frontier, between prediction accuracy and fairness. In theory, there is a lack of characterization of the Pareto frontier (optimal trade-off) between utility and fairness. Current works on the Pareto frontier, such as [37], apply tight inequalities based on the convexity of distance metrics to suggest the optimal trade-off coincide with the Wasserstein geodesic path. While such inequalities are tight for a broad type of metrics on the space of probability measures, they are *not tight* for the Wasserstein metric. Hence, the inequalities are not able to extend the mathematically rigorous Wasserstein barycenter characterization of the optimal fair learning outcome to a Pareto frontier. From a computational perspective, current methods, such as [37], apply interpolation between the inverses of the sensitive conditional cdf's (more specifically, interpolating the data points that share the same image under the sensitive conditional cdf's) to estimate the geodesics. In addition to the drawbacks mentioned above, the inverse of the cdf also does not come with an explicit form, which makes the computation of an interpolation between two cdf inverses even more cumbersome.

3. The post-processing nature of the characterization requires explicit or implicit sensitive information in the training and decision-making process. More specifically, in order to apply the barycenter characterization to find the optimal fair learning outcome or to make predictions to newly incoming data, one needs the following steps: (1) Estimate the conditional expectation and obtain its conditional distributions with respect to the sensitive information; (2) Find the Wasserstein barycenter of the sensitive conditionals of the conditional expectation estimation or the learning outcome; (3) Compute the optimal transport maps from each sensitive conditional to the barycenter; (4) Apply each transport map to the conditional with the matched sensitive information. Here, not only does the trained model still inherit unfairness, but it is also clear that sensitive information needs to be attached to both the dependent variable or incoming data and its learning outcome or prediction, until the very last post-processing step of finding the barycenter comes to the rescue. Hence, we say that

the characterization has a post-processing nature. As a result, the user needs access to the sensitive information of each individual incoming data at every step during the learning process. Such a strong access to sensitive information makes the supervised learning process vulnerable to attack and sensitive information leakage.

The post-processing nature of the characterization also suffers from the lack of flexibility in model selection, modification, and composition. For model selection and modification, a practitioner would have to perform the post-processing step for every model and every modification in order to compare the corresponding optimal fair learning outcomes. See Table 7 for more details on the additive computational cost of the post-processing approach compared to the one-time cost of the proposed pre-processing approach. For model composition, we consider the simple example $task_2 \circ task_1$ where $task_i, i \in \{1,2\}$ are trained supervised learning models. In practice, there is a good chance that $task_1$ and $task_2$ belong to different practitioners or organizations, denoted by practitioners 1 and 2, respectively. Therefore, to protect sensitive information from practitioner 2, practitioner 1 will perform the post-processing step to obtain a fair learning outcome and provide it as an input variable for the training task of practitioner 2. But unless $task_2$ needs no more input variables other than the dependent variables of $task_1$ (in that case, $task_1$ would be fair data representation design), still practitioner 2 needs full access to the sensitive variable attached to its input data, which includes the desensitized $task_1$ output and other input variables. Such attachment makes the post-processing step performed by practitioner 1 meaningless. Considering the recent development of decentralized learning in practice, such drawback in model composition makes a model-independent fair data representation more applicable than a post-processing solution.

4. Many of the current fair machine learning methods are proposed without utility guarantee or explainability. Such a lack of utility guarantee or explainability prevents the study of fair machine learning from practical use. For instance, Wells Fargo [45] concluded recently that current fair machine learning methods are black-box methods, and hence they hesitate to adopt fair machine learning techniques.

We provide a road map of the tools that we have developed in response to each of the listed challenges and how the present work combines all the tools to provide (exact solution and estimation of) the fair data representation at the Pareto frontier.

1. In response to the theoretical part of the first challenge, Lemma 3.1 in Section 3 provides a characterization (with explicit construction) of the exact solution to Problem 1 (the optimal fair L^2-objective learning). The result shows that the infimum loss value of Problem 1 can be nicely decomposed into two parts: (1) L^2 orthogonal projection loss and (2) independence projection loss. Also, the result now allows the data spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to be $[k]^d, \mathbb{N}^d, [0,l]^d$, or $\mathbb{R}^d$ for arbitrary dimension $d < \infty$.

   To address the challenge of computing the Wasserstein barycenter in high-dimensional data spaces [3], we propose a method that applies affine transport maps to find the *optimal affine estimation* of the post-processing optimal fair L^2-objective supervised learning outcome with an arbitrarily finite-dimensional dependent variable, which responds to the first challenge listed above. In particular, by restricting admissible transport maps to be affine and making a corresponding relaxation to the fairness constraint, we derive a relaxed version of Problem 1, stated as Problem 4. Applying the optimal affine transport maps [2], Definition 3.1 introduces the post-processing *pseudo-barycenter*, Lemma 3.2 shows the proposed pseudo-barycenter coincides with the true barycenter when the sensitive conditionals are Gaussian, and finally, Theorem 3.1 proves that the pseudo-barycenter is the optimal affine estimation of the true barycenter in the general conditional distribution case and provides the estimation error. Optimal affine transport and pseudo-barycenter have the advantage of computational efficiency, compared to the current methods, due to the explicit matrix form of the transport map and the nearly closed-form solution to the pseudo-barycenter.

The importance of optimal affine maps encompasses much more than a solution to the first challenge. The optimal affine maps together with McCann interpolation [32] help us in obtaining an explicit form of the geodesic path characterization of the Pareto frontier in Section 4. More importantly, Section 5 shows that optimal affine maps and the pseudo-barycenter are necessary tools to overcome the post-processing nature of the Wasserstein barycenter characterization by exploiting the linearity of conditional expectation and thereby generating optimal fair data representations.

2. In Section 4, we prove an exact characterization of the solution to Problem 2 (the optimal utility-parity trade-off or Pareto frontier) in response to the theoretical part of the second challenge. In particular, Theorem 4.1 shows that, when utility loss and disparity are quantified respectively by the L^2 distance (between the true outcome Y and the prediction $\hat{Y} = f(X, Z)$) and the average pairwise $\mathcal{W}_2$ distance among the sensitive conditionals of $\hat{Y}$, the optimal trade-off happens if and only if the conditionals of $\hat{Y}$ travel along the Wasserstein geodesic path from the conditionals of $\mathbb{E}(Y|X, Z)$ to their barycenter. Therefore, we say that the Pareto frontier is on the Wasserstein space. Corollary 4.1 then derives an explicit form of the Pareto optimal solution to Problem 2. The result is a natural extension to the post-processing Wasserstein barycenter characterization of the optimal fair learning outcome: the barycenter characterization coincides with the point at zero disparity on the Pareto frontier. Interestingly, our result shows that the Pareto frontier is linear.

   To solve the computational challenge of the geodesic path, Remark 4.1 applies McCann interpolation together with the optimal affine maps and the pseudo-barycenter to derive a computationally efficient (nearly) closed-form formula to estimate the Pareto frontier, which results in Algorithm 1.

3. In response to the third challenge, the present work proposes in Section 1.3 Problem 3 (optimal fair data representation problem), which makes the objective function and the fairness (statistical parity) constraint *model-independent* and therefore suitable for fair data representation design. More specifically, by applying the Minkowski inequality, we use an objective function to maximize the potential utility remaining in the data. On the other hand, a fair data representation should provide a fairness guarantee to arbitrary L^2-objective supervised learning models. Therefore, the present work proposes a pre-processing fairness constraint to guarantee fairness in the learning outcome of arbitrary L^2-objective models trained via the fair data representation.

In Section 5, Lemma 5.3 first provides a characterization of the exact solution to Problem 3 under a mild assumption. Next, Definition 5.2 and Definition 5.1 define the dependent and independent pseudo-barycenter, respectively. Then, similar to solving a relaxation of the post-processing characterization to obtain the optimal affine estimation, Theorem 5.1 proves that the dependent and independent pseudo-barycenter pair coincides with the true solution to the optimal fair data representation when the conditional data distributions are Gaussian, and Theorem 5.2 proves that the pseudo-barycenter pair forms the optimal affine estimation of the optimal fair data representation.

To derive (an estimation of) fair data representation at the Pareto frontier, Corollary 5.1 in Section 5.4 first provides a characterization of the Pareto frontier for conditional expectation on a fixed sigma-algebra. Finally, combining optimal affine map, pseudo-barycenter, together with a diagonal argument in Remark 5.4, we derive an estimation of the fair representation at the Pareto frontier, which results in Algorithm 1 and Algorithm 2.

Furthermore, in Section 7, experiments show that the proposed fair data representations preserve as large an amount of information (w.r.t. the L^2 objective) as the fairness constraint allows. Therefore, it provides a better and more flexible solution to fair learning compared to encoding-based data representations [13, 44], which encode the information of the original data into some binary feature variables designed to guarantee statistical parity for classification. Surprisingly, experiments also show that applying the pseudo-barycenter results in nearly zero utility loss compared to the post-processing barycenter characterization solution.

4. In addition to the provable utility guarantee resulting from the Pareto frontier, the proposed method also has a meaningful interpretation from a *datapoint-wise perspective* in how it achieves the statistical parity requirement: A data point of the optimal fair learning outcome is the Euclidean average of the optimally matched data points from each of the sensitive groups. Here, matching means partitioning the original data set into subsets consisting of one point from each sensitive group. Each subset is called a match. The points within a match are called matched points. Optimality in matching is equivalent to minimization of the expected variance within a randomly chosen match. Such expected (hence total) variance minimization enforces points with similar relative positions in their sensitive marginal to form a match. For example, assume that there are two sensitive conditionals $A = \{1 \text{ (low in A)}, 4 \text{ (high in A)}\}$ and $B = \{2 \text{ (low in B)}, 3 \text{ (high in B)}\}$, then the optimal matching is

$$\{\{1 \text{ (low in A)}, 2 \text{ (low in B)}\}, \{3 \text{ (high in B)}, 4 \text{ (high in A)}\}\}$$

to minimize the expected or total variance within the matches. The optimal matching in high-dimensional L^2 spaces shares the same geometric intuition with the simple example. That is, from a point-wise perspective, the optimal fair learning achieves statistical parity by first matching the points with similar relative positions in their sensitive groups and then representing the matched ones with their Euclidean average.

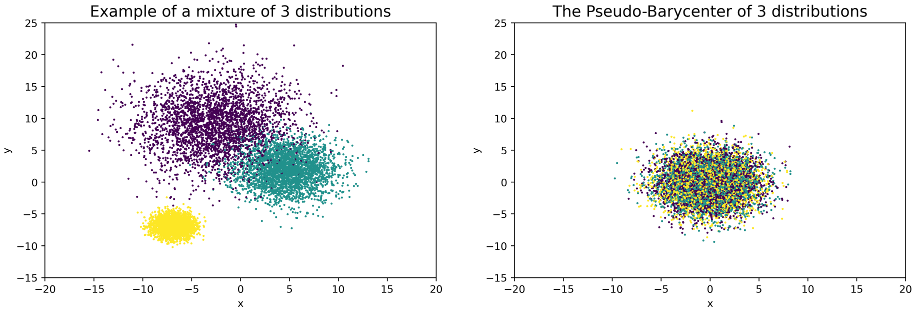


Figure 1: The left panel depicts three distributions, sampled from an isotropic Gaussian distribution with different first two moments. The right panel shows the pseudo-barycenter of the three sample distributions.

1.3 Fair Data Representations: From Theory to Practice

In this subsection, we derive a fairness objective function that is both theoretically tractable and practically appealing. This task is more involved than one initially might expect, and it sheds light on some subtleties of both the post-processing and the pre-processing approaches.

Before proceeding, we need some preparation. Let X, Y, and Z represent respectively the independent, dependent, and sensitive random variable, with the same underlying probability space $(\Omega, \Sigma, \mathbb{P})$. We use the term 'random variables' to denote random vectors with an arbitrary but finite dimension. That is, $S : \Omega \to \mathcal{S}$ where $\mathcal{S} \in \{[k_{\mathcal{S}}]^{d_{\mathcal{S}}}, \mathbb{N}^{d_{\mathcal{S}}}, [0, l_{\mathcal{S}}]^{d_{\mathcal{S}}}, \mathbb{R}^{d_{\mathcal{S}}}\}$ with $k_{\mathcal{S}} \in \mathbb{N}, l_{\mathcal{S}} \in \mathbb{R}$ and $d_{\mathcal{S}} < \infty$ for $S \in \{X, Y, Z\}$.

It follows from [18, 24] that the optimal fair regression outcome can be characterized by the Wasserstein barycenter. In Lemma 3.1 we will generalize their result from regression to all functions in $L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$, which shows that the optimal fair L^2-objective supervised learning outcome can be characterized by solutions to Problem 1:

$$\inf_{f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})} \{||Y - f(X, Z)||_2^2 : f(X, Z) \perp Z\} \tag{4}$$

The utility loss is quantified by L^2-norm: $||Y - f(X, Z)||_2^2 = \int_\Omega ||Y - f(X, Z)||^2 d\mathbb{P}$, where $|| \cdot ||$ is the Euclidean norm. The constraint $f(X, Z) \perp Z$ guarantees that the final result satisfies statistical parity and, therefore, is fair.

Since it follows from L^2 orthogonal decomposition that

$$||Y - f(X, Z)||_2^2 = ||Y - \mathbb{E}(Y|X, Z)||_2^2 + ||\mathbb{E}(Y|X, Z) - f(X, Z)||_2^2 \tag{5}$$

and only the second term on the right hand side depends on the choice of $f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$, we conclude that (1) is equivalent to

$$\inf_{f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})} \{||\mathbb{E}(Y|X, Z) - f(X, Z)||_2^2 : f(X, Z) \perp Z\}. \tag{6}$$

It turns out—see Lemma 3.1—that the solution to (6) is exactly the Wasserstein barycenter. Therefore, we say that the optimal fair L^2-objective supervised learning outcome is characterized by the Wasserstein barycenter. But notice that the Wasserstein barycenter characterization (6) assumes knowledge of the learning outcome $\mathbb{E}(Y|X, Z)$. That is, if practitioners apply the characterization to estimate the optimal learning outcome, it is necessary to obtain an estimator of $\mathbb{E}(Y|X, Z)$ via supervised learning before solving the post-processing rescue step (6). Therefore, we say that the characterization has a post-processing nature and hence call it a post-processing characterization.

Now, notice that the estimator of $\mathbb{E}(Y|X,Z)$ is obtained via the training process

$$\inf_{f\in\mathcal{F}}\{||Y-f(X,Z)||_2^2\}, \tag{7}$$

where the admissible function set $\mathcal{F}$ depends on the choice of supervised learning models. Denote the estimator by $f'(X,Z)$. Then in practice (6) becomes

$$\inf_{f\in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})}\{||f'(X,Z)-f(X,Z)||_2^2 : f(X,Z)\perp Z\}. \tag{8}$$

That is, the application of the post-processing characterization is model-dependent. The fundamental reason for model dependence is that (1) is optimizing over all L^2 functions while in practice it is necessary to reduce the admissible set from L^2 to some $\mathcal{F}$ which depends on the choice of the model. As a result, the optimizer is necessarily dependent on the choice of the model. Therefore, the constrained optimization (1) and its characterization are not suitable for our ultimate goal of deriving a model-independent pre-processing approach to the optimal fair learning outcome. The present work proposes a different constrained optimization problem that characterizes the optimal fair data representation for all L^2-objective supervised learning models.

To make a constraint optimization problem suitable for fair data representation design, we require both the objective function and the fairness constraint to be model-independent. Furthermore, the data representation design objective and the training objective given the data representation have to be consistent in the following sense: the better training and testing result on the fair data representation leads to less L^2-fitting error with respect to the true data.

We now derive an objective function that is suitable for fair data representation design purpose. To start, notice that our goal is to generate a synthetic data representation $(\tilde{X},\tilde{Y})$, a deformation of (X,Y), via which any L^2-objective model that is trained by

$$\inf_{f\in\mathcal{F}}||\tilde{Y}-f(\tilde{X})||_2^2 \tag{9}$$

would result in (an estimation of) the optimal fair learning outcome. In the rest of this paper we denote the solution to (9) by $f_{\tilde{Y}}$.

Also, because conditional expectation is an orthogonal projection operator on L^2-space, we obtain the following orthogonal decomposition of the objective in (9):

$$||\tilde{Y}-f(\tilde{X})||_2^2 = ||\tilde{Y}-\mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 + ||\mathbb{E}(\tilde{Y}|\tilde{X})-f(\tilde{X})||_2^2. \tag{10}$$

Only the second term on the right hand side depends on the choice of $f\in\mathcal{F}$, hence the training step objective (9) is equivalent to the following:

$$\inf_{f\in\mathcal{F}}||\mathbb{E}(\tilde{Y}|\tilde{X})-f(\tilde{X})||_2^2. \tag{11}$$

Thus, the solution to (11) is also $f_{\tilde{Y}}$, which depends on the choice of $\mathcal{F}$.

The key observation is that, given a data representation $(\tilde{X}, \tilde{Y})$, (11) is the objective that practitioners try to achieve via model selection, modification, and parameter turning. Furthermore, it follows from the triangle or Minkowski inequality that

$$\underbrace{||Y - f_{\tilde{Y}}(\tilde{X})||_2}_{\text{total utility loss}} \leq \underbrace{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2}_{\text{data representation utility loss}} + \underbrace{||\mathbb{E}(\tilde{Y}|\tilde{X}) - f_{\tilde{Y}}(\tilde{X})||_2}_{\text{learning utility loss}}. \tag{12}$$

The second term on the right-hand side is the target of a supervised learning task which should be left to practitioners. Thus, the natural choice of the model-independent objective of the optimal fair synthetic data design is to minimize the first term:

$$\inf_{(\tilde{X},\tilde{Y})\in\mathcal{D}} ||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2, \tag{13}$$

where $\mathcal{D}$ is some admissible set of deformed versions of the original data (X, Y) that we define later. Intuitively, the loss function can be interpreted as the potential utility sacrifice resulting from deforming (X, Y) to $(\tilde{X}, \tilde{Y})$ for L^2-objective supervised learning, while leaving the task of minimizing the second term on the right-hand side to practitioners via model selection, modification, or parameter tuning.

Next, we derive a fairness constraint for synthetic data design purposes. That is, the goal is to design $(\tilde{X}, \tilde{Y})$ such that $f_{\tilde{Y}}(\tilde{X}) \perp Z$ for any admissible function set $\mathcal{F} \subset L^2(\mathcal{X}, \mathcal{Y})$. The flexibility of model choice becomes important due to the increasing complexity of models in practice nowadays, such as neural networks. The key observation here is that, due to the potential dependence of $f_{\tilde{Y}}$ on Z, one needs to look at both models that use merely measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ and more complicated models consisting of Z-dependent measurable functions:

1 For measurable functions from $\mathcal{X}$ to $\mathcal{Y}$, if we require $\tilde{X} \perp Z$, then it follows that for any $f : \mathcal{X} \to \mathcal{Y}$, it is guaranteed that $f(\tilde{X}) \perp Z$. Hence, we require $\tilde{X} \perp Z$ to prevent models from exploiting sensitive information from the independent variables.

2 For advanced or adversarial models that use Z-dependent functions from $\mathcal{X} \times \mathcal{Z}$ to $\mathcal{Y}$, the trained model f_Y could still depend on Z because Y and Z are not independent. For example, consider the extreme case where $Y = kZ, k \in \mathbb{R}$ and a perfect model results in $\mathbb{E}(kZ|\tilde{X}, Z) = kZ$ which fully depends on Z even if we require $\tilde{X} \perp Z$. Therefore, we also require $f_{\tilde{Y}}(\tilde{X}, Z) \perp Z$ to prevent such a model from exploiting sensitive information from the dependent variables.

But notice that the second requirement leads us back to the post-processing nature of fairness constraints as in (8). For fair data representation design purposes, it is necessary to keep the constraint model-independent. Therefore, instead of enforcing $f_{\tilde{Y}}(\tilde{X}, Z) \perp Z$, the present work requires $\mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z$ for the following two reasons: (1) Under the modified constraint $\mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z$, the better $f_{\tilde{Y}}(\tilde{X}, Z)$ estimates $\mathbb{E}(\tilde{Y}|\tilde{X}, Z)$, the more independent of Z becomes $f_{\tilde{Y}}(\tilde{X}, Z)$. Such alignment between training objective and fairness makes the modification a natural choice under the assumption that the goal of L^2-objective (adversarial) supervised learning tasks is to minimize $||\mathbb{E}(\tilde{Y}|\tilde{X}, Z) - f_{\tilde{Y}}(\tilde{X}, Z)||_2^2$, which is equivalent to minimizing $||\tilde{Y} - f_{\tilde{Y}}(\tilde{X}, Z)||_2^2$. (2) Since a supervised learning model with poor prediction accuracy already results in severe unfairness, the dependence on sensitive information is of less concern when designing a fair data representation.

Based on the fairness requirement for both measurable functions on merely $\mathcal{X}$ and Z-dependent functions, a natural choice of (pre-processing) statistical parity constraint for data representation has the following form:

$$\tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z. \tag{14}$$

It guarantees: (1) statistical parity for any model that uses only a deterministic function and any model that results in a perfect estimation of $\mathbb{E}(\tilde{Y}|\tilde{X})$; (2) the better $f_{\tilde{Y}}(\tilde{X}, Z)$ estimates $\mathbb{E}(\tilde{Y}|\tilde{X}, Z)$, the more independent $f_{\tilde{Y}}(\tilde{X}, Z)$ becomes of Z.

While the fairness constraint (14) is not the only choice, it does balance utility and fairness. The following remark discusses two alternative fairness constraint choices, which are more polarized in optimizing utility or fairness.

Remark 1.3 (Alternative fair data representation constraints) *There are two alternative choices of fairness constraints that are valuable in practice:*

1 $\tilde{X} \perp Z$: the weaker constraint guarantees any model using merely a deterministic function, even if sub-optimal, to result in statistical parity. But it does not protect Z from advanced models, which exploit the dependence of Y on Z and apply Z-dependent functions. Therefore, $\tilde{X} \perp Z$ provides more utility but less sensitive information protection, compared to our choice.

2 $(\tilde{X}, \tilde{Y}) \perp Z$: the stronger constraint guarantees statistical parity in the learning outcome of any supervised learning model, even for those that adopt Z-dependent functions and are suboptimal. But it sacrifices more utility. This stronger constraint is particularly useful in practice when one does not know which variables are dependent and which ones are independent.

Our choice is a compromise of the two alternatives in terms of balancing utility sacrifice and protecting sensitive information. Furthermore, simple modifications of our analysis and algorithm would solve the two alternatives because they are essentially simplified versions of our choice. Hence, the present work targets (14).

Finally, combining the objective and constraint for synthetic data design, we aim to solve Problem 3:

$$\inf_{(\tilde{X}, \tilde{Y}) \in \mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : \tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z\}. \tag{15}$$

The solution provides a fair data representation via which the trained L^2-objective supervised learning models become estimations of the optimal fair conditional expectation.

Compared to the original constrained optimization problem (1) which results in the post-processing nature of its barycenter characterization (6), the proposed constrained optimization problem (3) has the following advantages by design:

1 It provides a fairness guarantee for arbitrary L^2-objective models.

2 The model-independence together with the alignment between training objective and fairness enables practitioners to enjoy flexibility in model selection, modification, and parameter tuning on the fair data representation.

3 The fair data representation approach has more applicable models than the post-processing approach. See Remark 1.4 below for a detailed explanation of two different interpretations of L^2-objective models.

In the following remark, we explain the different interpretations of L^2-objective models in the post-processing and pre-processing approaches.

Remark 1.4 (Interpretation of L^2-objective models) *For the post-processing approach, it follows from* (6) *and* (8) *that the barycenter characterization works only if the supervised learning model comes with an objective function in explicit L^2-form. For the proposed pre-processing approach, the applicable L^2-objective models include all the models that aim to estimate the conditional expectation. In particular, it follows from* (12) *and* (13) *that the proposed fair data representation works for any supervised learning model that aims to estimate conditional expectation or conditional probability, even though some of them do not come with an explicit objective function in L^2-form. For example, all classification models share the goal of estimating the conditional probability of $\{Y = 1\}$ given an observation of $\{X = x\}$, which is $\mathbb{E}(\mathbb{1}_{Y=1}|X = x)$. Therefore, the resulting synthetic data can be used for any classification model, even models such as logistic regression and random forest that do not have L^2-based objective functions.*

1.4 Setting and Notation

In the rest of the work, $\mathcal{L}(X) = \mathbb{P} \circ X^{-1} : \mathcal{B}_{\mathcal{X}} \to [0, 1]$ denotes the distribution or law of X, which is a function that assigns each event in the Borel sigma-algebra, $\mathcal{B}_{\mathcal{X}}$, a probability. Let $\lambda := \mathcal{L}(Z)$ denote the law of the sensitive random variable to simplify notation. To remove sensitive information Z, the method we propose is to find a set of maps $T_x := \{T_x(\cdot, z)\}_z$ such that $T_x(\cdot, z) : \mathcal{X} \to \mathcal{X}$ pushes the conditional (on $\{Z = z\}$) distribution (see the definition of conditional distribution $\mathcal{L}(X_z)$ below) forward to a common probability measure $\mathcal{L}(\tilde{X})$ for λ-a.e. $z \in \mathcal{Z}$. Also, when restricting T to be a linear map or a matrix, we use $T \succ 0$ to denote T is positive definite, and $||T||_F$ to denote its Frobenius norm.

Given a measurable map $T : \mathcal{X} \to \mathcal{X}$ and a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, $T_\sharp\mu$ denotes the push-forward probability measure that is defined as the following: for any event, A, in the Borel sigma-algebra, $\mathcal{B}_{\mathcal{X}}$, $T_\sharp\mu(A) := \mu(T^{-1}(A))$. In the rest of the paper, we often say T pushes μ forward to $T_\sharp\mu$.

The conditional distributions $\{\mathcal{L}(X_z)\}_z$ are defined uniquely λ-a.e. by the disintegration theorem [36, Box 2.2]. Hence, $z \to \mathcal{L}(X_z)$ is Borel measurable and, for all Borel measurable sets $E \in \mathcal{B}_{\mathcal{X}}$, $\mathbb{P}(E) = \int_{\mathcal{X}} \mathbb{P}(X_z^{-1}(E))d\lambda(z)$. The application of the disintegration theorem aims to allow $\mathcal{Z}$ to be uncountably infinite, such as the real line or the real vector space. In the practical case of a finite data set, when the data set (X, Z) is $\{(x_i, z_i)\}_{i\in[N]}$, for each $z \in \mathcal{Z}$, the empirical conditional random variable (with uniform distribution) is defined as follows:

$$X_z := \{x_i : (x_i, z_i) \in (X, Z), z_i = z\}.$$

Therefore, on the product data space $\mathcal{X} \times \mathcal{Z}$ with a joint distribution, the law of the random variable or vector X_z is the conditional distribution on $\{Z = z\}$.

The present work often assumes the conditionals $\{\mathcal{L}(X_z)\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{X})$. Here, $\mathcal{P}_{2,ac}(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$ that have finite second moments and are absolutely continuous with respect to the Lebesgue measure. The finite second moment assumption guarantees the Wasserstein distance to be well-defined without being infinite. The absolute continuity assumption guarantees the existence of their Wasserstein barycenter (See Definition 2.3) and the respective (almost surely invertible) optimal transport maps that map them to the barycenter. The present work denotes the barycenter by $\overline{\mathcal{L}(X_z)}$ or $\overline{\mathcal{L}(X)}$ interchangeably, and denotes the optimal transport map that pushes $\mathcal{L}(X_z)$ to $\overline{\mathcal{L}(X)}$ by T_z or $T(\cdot, z)$.

To simplify notation and proof, we define $\bar{X}$ to be the random variable that satisfies the following: for λ-a.e. $z \in \mathcal{Z}$,

$$\bar{X}_z = T_z(X_z). \tag{16}$$

In other words, the couple $(X_z, \bar{X}_z)$ is a coupling of $(\mathcal{L}(X_z), \overline{\mathcal{L}(X)})$ and satisfies:

$$||X_z - \bar{X}_z||_2^2 = \mathcal{W}_2^2(\mathcal{L}(X_z), \overline{\mathcal{L}(X)}) \tag{17}$$

for λ-a.e. $z \in \mathcal{Z}$. We refer interested readers to [40, 41] for more details on the assumption of $\mathcal{P}_{2,ac}(\mathcal{X})$ and the coupling of measures. In the rest of the paper, we call $\bar{X}$ the Wasserstein barycenter of $\{X_z\}_z$.

In solving the post-processing characterization, with the assumption of $\mathbb{E}(Y|X, Z)$, one first finds the Wasserstein barycenter of $\{\mathcal{L}(\mathbb{E}(Y|X, Z)_z)\}_z$, denoted by $\overline{\mathcal{L}(\mathbb{E}(Y|X, Z)_z)})$. Here, $\mathbb{E}(Y|X, Z)_z$ denotes the conditional of $\mathbb{E}(Y|X, Z)$ on $\{Z = z\}$ for λ-a.e. $z \in \mathcal{Z}$. Then one applies the optimal transport map $T(\cdot, z) : \mathcal{Y} \to \mathcal{Y}$ which pushes $\mathbb{E}(Y|X, Z)_z$ forward to $\overline{\mathbb{E}(Y|X, Z)}_z$ for λ-a.e. $z \in \mathcal{Z}$.

In solving the pre-processing characterization, one has two different optimal transport maps to deform X and Y. For the dependent variable, we define $T_y = \{T_y(\cdot, z)\}_z$, $\mathcal{L}(Y_z)$, and $\mathcal{L}(\tilde{Y})$ analogously, but require merely the agreement of $\mathcal{L}(\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_z)$ for λ-a.e. $z \in \mathcal{Z}$. The λ-a.e. agreement of $\mathcal{L}(\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_z)$ means that the laws of the random variables or vectors $\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_z$ are equal, except for some z on a λ-null set on $\mathcal{Z}$. In other words, on the Borel measurable space $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, for any set B in the Borel sigma-algebra $\mathcal{B}_{\mathcal{Y}}$, we have $\mathbb{P} \circ [\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_{z_1}]^{-1}(B) = \mathbb{P} \circ [\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_{z_2}]^{-1}(B)$ for all $z_1, z_2 \in \mathcal{Z}$, except on a set $N \subset \mathcal{Z}$ such that $\lambda(N) = 0$.

Therefore, by generating and applying (T_x, T_y) to the data, we achieve $\mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z$, i.e. *statistical parity*, due to the enforced λ-a.e. agreement of $\mathcal{L}(\mathbb{E}(\tilde{Y}|\tilde{X}, Z)_z))$. Combining the application of deformation maps and (3), we obtain the fair data representation optimization problem

$$\inf_{(\tilde{X},\tilde{Y})\in\mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : \tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z\} \tag{18}$$

with the admissible set $\mathcal{D}$ is defined as

$$\mathcal{D} := \{(\tilde{X}, \tilde{Y}) : \tilde{X} = T_x(X, Z), \tilde{Y} = T_y(Y, Z)\}, \tag{19}$$

Here, $T_x(\cdot, z) : \mathcal{X} \to \mathcal{X}$ and $T_y(\cdot, z) : \mathcal{Y} \to \mathcal{Y}$ are Borel measurable maps. We denote the set of admissible $\tilde{X}$ and $\tilde{Y}$ by $\mathcal{D}|_{\mathcal{X}}$ and $\mathcal{D}|_{\mathcal{Y}}$, respectively. The reason underlying the definition of $\mathcal{D}$ is that the fair data should still has its foundation from the real data, albeit suitably "deformed".

1.5 Paper Organization

The rest of the paper is organized as follows: Section 2 reviews the tools in optimal transport that are needed to derive results in the present work: Wasserstein space, Wasserstein barycenter, and optimal affine transport within a location-scale family. Section 3 first generalizes the current barycenter characterization of optimal regression to optimal L^2-objective supervised learning, then defines pseudo-barycenter, and proves pseudo-barycenter is the optimal affine estimation of the true barycenter. Section 4 is concerned with both the theoretical characterization and an explicit formula of the Pareto frontier on the Wasserstein space. Section 5 studies the exact solution to the optimal data representation and the optimal affine estimation of the exact solution. Section 6 proposes an algorithm based on the theoretical results in the previous sections. Section 7 provides an extensive numerical study regarding the application of the pseudo-barycenter and the optimal affine maps to (1) the estimation of optimal fair learning outcome compared to the known fair machine learning techniques on different learning models; and (2) Pareto frontier estimation for different disparity definitions.

2. Preliminaries on Optimal Transport

In this section, we review the theoretical results on optimal transport and the Wasserstein barycenter that are important for the development of the main theoretical results on efficient algorithm design, Wasserstein geodesic characterization of the Pareto frontier, and the pre-processing approach resulting in the optimal fair data representation. For our purposes, we focus on $\mathbb{R}^d$. We refer readers who are interested in more generalized versions, e.g. on compact Riemannian manifolds, to for example [30].

2.1 General Distribution Case

Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, which is the set of all probability measures on $\mathbb{R}^d$, Monge asked for an optimal transportation map $T_{\mu\nu} : \mathbb{R}^d \to \mathbb{R}^d$ that solves

$$\inf_{T_\sharp \mu = \nu} \Big\{ \int_{\mathbb{R}^d} ||x - T(x)||^2 d\mu \Big\} \tag{20}$$

Here, $|| \cdot ||$ denotes the Euclidean norm on $\mathbb{R}^d$. The problem remained open until Brenier showed that Monge's problem coincides with Kantorovich's relaxed version:

$$\inf_{\gamma \in \prod(\mu,\nu)} \Big\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x_1 - x_2||^2 d\gamma(x_1, x_2) \Big\} \tag{21}$$

and admits a unique solution provided $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. Here, $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ denotes the space of probability measures on $\mathbb{R}^d$ that have finite first two moments and are absolutely continuous w.r.t. (with respect to) the Lebesgue measure. That is, the optimal solution to (21) has the form: $\gamma = (Id, T_{\mu\nu})_\sharp \mu$, where $T_{\mu\nu}$ solves (20). Here, $\prod(\mu, \nu)$ denotes all the probability measures on $(\mathbb{R}^{2d}, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d))$ such that the marginals are μ and ν. The relaxed problem is easy to solve due to the weak* compactness of $\prod(\mu,\nu)$. We refer interested readers to [40, 41] for more detailed existence and uniqueness results.

Remark 2.1 *The uniqueness is in the weak sense for γ and μ-a.e. for $T_{\mu\nu}$.*

Kantorovich's problem provides a certain kind of "distance" on $\mathcal{P}(\mathbb{R}^d)$ except for the possibility of being infinite.

Definition 2.1 (Wasserstein distance[2]) *Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathcal{W}_2(\mu,\nu) := \Big(\inf_{\gamma \in \prod(\mu,\nu)} \Big\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x_1 - x_2||^2 d\gamma(x_1,x_2) \Big\}. \Big)^{\frac{1}{2}} \tag{22}$$

It is not hard to verify that the Wasserstein distance defined above satisfies the axioms of a metric except for finiteness of $\mathcal{W}_2(\mu,\nu)$ for arbitrary $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. In order to guarantee finiteness, one needs to put more restrictions on the set of all probability measures:

Definition 2.2 (Wasserstein space) *Define $\mathcal{W}_2$ as above and*

$$\mathcal{P}_2(\mathbb{R}^d) := \Big\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} ||x||^2 d\mu < \infty \Big\}. \tag{23}$$

The couple $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called Wasserstein space.

The Wasserstein space has gained increasing popularity in image processing, economics [22, 15], and machine learning in recent years due to its useful properties such as polishness (of the space) and robustness (w.r.t. perturbation on the marginal probability measures and hence on sampling).

Since the Wasserstein space is a metric space, the Fréchet mean on the space is well-defined and it is called the Wasserstein barycenter in the optimal transport literature.

Definition 2.3 (Wassserstein barycenter [2]) *Given $\{\mu_z\}_{z\in\mathcal{Z}} \subset (\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ for some index set $\mathcal{Z}$, the barycenter of $\{\mu_z\}_z$ is the Fréchet mean of the set on $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$. That is, $\bar{\mu}$ is the solution to*

$$\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \Big\{ \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) d\lambda(z) \Big\}, \tag{24}$$

where $\bar{\mu}$ denotes the Fréchet mean or barycenter.

Here, for our purpose, we focus on the case where the index set $\mathcal{Z} \in \{[k], \mathbb{N}, [0,1], \mathbb{R}^n\}$.

Next, we look at optimal transport and the barycenter problem from the perspective of optimal coupling. The goal is to show that the multi-marginal coupling problem is equivalent to the Wasserstein barycenter problem. The equivalence is an essential tool in proving our result in optimal affine transport, the optimality of the pseudo-barycenter, and the geodesic characterization of the Pareto frontier.

[2] Throughout this paper we work with the Wasserstein-2 distance, and thus simply call it the Wasserstein distance.

First, notice that Kantorovich's problem is in fact a 2-marginal coupling problem: Let X_1, X_2 be the random variable satisfy $\mathcal{L}(X_1) = \mu, \mathcal{L}(X_2) = \nu$, the problem looks for a γ with marginals being μ, ν that minimizes $\mathbb{E}_\gamma ||X_1 - X_2||^2$. It follows naturally by the existence and uniqueness result of the optimal transport map (also known as Brenier's map) [11], that the Wasserstein distance admits the form in the classic probability language:

$$\mathcal{W}_2(\mu, \nu) = (\mathbb{E}_\mu ||X_1 - T(X_1)||^2)^{\frac{1}{2}}, \tag{25}$$

where T is the optimal transport map that pushes $\mu = \mathcal{L}(X_1)$ forward to $\nu = \mathcal{L}(X_2)$.

More recent work in mathematics [30, 34] and economics [15, 22] has generalized the Kantorovich problem to the multi-marginal coupling problem:

$$\inf_{\gamma \in \prod(\{\mu_z\}_{z\in\mathcal{Z}})} \big\{\mathbb{E}_\gamma(\int_{\mathcal{Z}^2} ||X_{z_1} - X_{z_2}||^2 d\lambda(z_1) d\lambda(z_2))\big\}, \tag{26}$$

where $\prod(\{\mu_z\}_{z\in\mathcal{Z}})$ denotes all the Borel probability measures on $(\mathbb{R}^d)^{|\mathcal{Z}|}$ with marginals being $\mu_z = \mathcal{L}(X_z) \in \mathcal{P}(\mathbb{R}^d)$ λ-a.e.. Hence, one can consider $\lambda \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$. It can be shown that the above is equivalent to the following:

$$\sup_{\gamma \in \prod(\{\mu_z\}_{z\in\mathcal{Z}})} \big\{\mathbb{E}_\gamma(|| \int_{\mathcal{Z}} X_z d\lambda(z)||^2)\big\} \tag{27}$$

Remark 2.2 (Justification for the name of marginals) *Since $\{X_z\}_z$ are the marginals for the admissible couplings in* (26)*, with the equivalence between the multi-marginal coupling and Wasserstein barycenter (see Remark 2.3 below) in mind, we often call $\{X_z\}_z$ and $\{\mathcal{L}(X_z)\}_z$ the sensitive marginals, even though they are also the conditional random variables and distributions constructed by disintegration.*

Intuitively, (27) tends to find a family of random variables parametrized by z with fixed marginals μ_z such that the variance of the matched (by γ) group average is maximized. For readers who are more familiar with stochastic processes, consider $z = t$ as a time variable, then X_t is a stochastic process with fixed time marginals, and (27) tends to find a way (γ) to group the fixed marginals into trajectories so that the variance of the trajectory-wise (sample path) average is maximized. (Hence, the expected variance within a randomly chosen sample path is minimized.)

As shown in [2, 34], the above multi-marginal problem is equivalent to the barycenter problem:

Remark 2.3 (Equivalence between multi-marginal coupling and barycenter) *Assume $\{\mu_z\}_z$ are absolutely continuous w.r.t. the Lebesgue measure and let γ^* and $\bar{\mu}$ be the solution to* (27) *and* (24)*, respectively. It follows that $\bar{\mu} = \gamma^* \circ T^{-1}$ where $T(\{x_z\}_z) := \int_{\mathcal{Z}} x_z d\lambda(z)$.*

The importance of this equivalence is twofold:

1 It is the key to proving the non-degenerate Gaussianity of the Wasserstein barycenter of non-degenerate Gaussian marginal distributions;

2 It provides technical support for the interpretation (Section 1.3 point 4) of how the Wasserstein barycenter solves data-related fairness issues on a point-wise scale.

Therefore, we generalize the equivalence to the case where $\mathcal{Z}$ is a Polish space, which is a metric space that is separable and complete. In particular, $[k]^d, [0,l]^d, \mathbb{N}^d, \mathbb{R}^d$ mentioned above are all examples of Polish spaces. This generalization is important for our purpose as it provides a theoretical foundation for removing Z in the form of random vectors.

Now, the following result provides the existence and uniqueness result of the barycenter problem that is suitable for our purpose.

Theorem 2.1 (Existence and uniqueness of barycenter [31](Theorem 2 and Proposition 6))

Assume that $\mathcal{Z}$ is a Polish space and that $\lambda := \mathbb{P} \circ Z^{-1}$ satisfies $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \nu) d\lambda(z) < \infty$ for some $\nu \in \mathcal{P}_2(\mathcal{X})$ (hence, for all $\nu \in \mathcal{P}_2(\mathcal{X})$). Then the following properties hold:

1 There exists a barycenter of $\{\mu_z\}_{z \in \mathcal{Z}}$ w.r.t. λ.

2 If, in addition, $\lambda(\{z : \mu_z \in \mathcal{P}_{ac}(\mathcal{X})\}) > 0$, then the barycenter is unique.

Remark 2.4 (Applicability of assumptions in Theorem 2.1) *The assumption that $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \nu) d\lambda(z) < \infty$ in the above result is satisfied in our application to the optimal fair learning outcome or data representation: When generating the optimal transport maps $\{T_z\}_z$, the training set has a finite number of data and hence finite different values of z in the discrete case or after discretization in the continuous case. Therefore, since $\{\mu_z\}_z \subset \mathcal{P}_2(\mathcal{X})$, pick a value z_0 that is in the training set, we have that $\mathcal{W}_2^2(\mu_z, \mu_{z_0})$ are essentially (w.r.t. λ) uniformly bounded. That implies $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu_{z_0}) d\lambda(z) < \infty$.*

Now, we have the theoretical results that are needed to prove the main results, except for the McCann interpolation, which will be introduced in Section 4. The next step is to develop a computationally efficient method to compute (an estimation of) the Wasserstein barycenter, (the McCann interpolation of) optimal transport maps, and thereby the optimal fair model and Pareto frontier. More specifically, we focus on positive definite affine optimal transport maps.

2.2 Rigid Translation

Before deriving our main result on optimal positive definite affine maps, we first study the case where admissible maps are restricted to the set of rigid translations. The following property of rigid translations makes our results on the optimal affine maps simpler: we can assume, without loss of generality, that the first moments of the marginal measures are zero: $m_{X_z} := \mathbb{E}(X_z) = 0$ and $m_{Y_z} := \mathbb{E}(Y_z) = 0$.

Lemma 2.1 *Let $\mu, \nu \in \mathcal{P}_2$, $m_\mu := \int x d\mu(x)$, and $m_\nu := \int x d\nu(x)$. Also, let μ', ν' be the centered versions of μ, ν, respectively. It follows that*

$$\mathcal{W}_2^2(\mu, \nu) = \mathcal{W}_2^2(\mu', \nu') + ||m_\mu - m_\nu||^2. \tag{28}$$

Proof See Appendix A. ■

Notice that the above result allows us to assume measures to have vanishing first moments when deriving the optimal transport maps. Indeed, if $T_{\mu'\nu'}$ is the Brenier's map between μ' and ν', then $T_{\mu\nu} := T_{+m_\nu} \circ T_{\mu'\nu'} \circ T_{-m_\mu}$ is the optimal transport map between μ and ν. Here, $T_{+m_\nu}(x) := x + m_\nu$ and T_{-m_μ} are defined analogously.

In the rest of Section 2, we assume without loss of generality that the first moments of the measures are all equal to zero.

2.3 Location-Scale Case and Optimal Affine Transport

A sufficient condition for Brenier's maps to be positive definite affine is to require a certain "similarity" between the marginal data distributions. One natural choice is to assume $\{Y_z\}_z$ and $\{X_z\}_z$ to be non-degenerate Gaussian vector λ-a.e.. As shown in [4], the assumptions of Gaussian vector can easily be generalized to a *location-scale family*. In the definition below, $\mathcal{S}^d_{++}$ denotes the set of all $d \times d$ positive definite matrices.

The generalization from Gaussian to location-scale families is important for the main result in the next section, where we consider computationally efficient solutions to a relaxation of the Wasserstein barycenter problem in the case of general marginal distributions.

Definition 2.4 (Location-Scale Family) *For any $\mathcal{L}(X_0) \in \mathcal{P}(\mathbb{R}^d)$, define*

$$\mathcal{F}(\mathcal{L}(X_0)) := \left\{\mathcal{L}(AX_0 + m) : A \in \mathcal{S}^d_{++}, m \in \mathbb{R}^d\right\}. \tag{29}$$

The set $\mathcal{F}(\mathcal{L}(X_0))$ is called a location-scale family characterized by $\mathcal{L}(X_0)$.

In other words, under the assumption of vanishing first moments, the random variables that share laws in the same location-scale family can be transformed into each other by a positive definite linear transformation.

In [4] it is shown that Brenier's map between two probability measures, each having a vanishing first moment, within the same location-scale family is linear and has a closed form.

Lemma 2.2 (Optimal affine map) *If $\mu, \nu \in \mathcal{F}(\mathcal{L}(X_0))$ for some X_0 such that $m_\mu = m_\nu = 0$, then the Brenier's map that pushes μ forward to ν is given by:*

$$T_{\mu\nu} = \Sigma_\mu^{-\frac{1}{2}}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}} \tag{30}$$

where $\Sigma_\mu := \int xx^T d\mu$ and $\Sigma_\nu := \int xx^T d\nu$.

Proof See, for example, Theorem 2.3 in [4]. ■

Remark 2.5 *The optimal affine map is also the midpoint of the geodesic path joining Σ_μ^{-1} and Σ_ν on the manifold of positive definite matrices. We refer interested readers to, for example, Chapter 6.1 in [9] for more details.*

Now, back to the barycenter problem. It follows from Lemma 2.2 that, if one assumes that all the marginals belong to the same location-scale family, then the barycenter also belongs to the family and a nearly closed-form solution to the barycenter is available.

Lemma 2.3 (Barycenter in the location-scale case) *Assume $\{\mu_z\}_z$ belong to the same location-scale family $\mathcal{F}(P_0)$ and satisfy $m_{\mu_z} = 0, \Sigma_{\mu_z} \succ 0, \lambda-a.e.$, then there exists a unique solution, denoted by $\bar{\mu}$, to* (24)*. Moreover, $\bar{\mu}$ also belongs to $\mathcal{F}(P_0)$ and is characterized by $m_{\bar{\mu}} = 0$ and $\Sigma_{\bar{\mu}} = \Sigma$ where Σ is the unique solution to the following equation:*

$$\int_{\mathcal{Z}} (\Sigma^{\frac{1}{2}} \Sigma_{\mu_z} \Sigma^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) = \Sigma, \tag{31}$$

where Σ_{μ_z} is the second moment of $\mu_z, \forall z \in \mathcal{Z}$.

Proof See Appendix A. ■

In the case where $m_{\mu_z} \neq 0$, it follows from Lemma 2.1 that

$$\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) d\lambda(z) = \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \mu') d\lambda(z) + \int_{\mathcal{Z}} ||m_{\mu_z} - m_{\mu}||^2 d\lambda(z)$$

where μ' denotes the centered version of μ. By Lemma 2.3, we know the first term on the right is minimized at $\bar{\mu}' \sim \mathcal{N}(0, \Sigma_{\bar{\mu}})$. Also, the second term on the right is minimized at the Fréchet mean with Euclidean metric, which is equal to the expectation. That is, $m_{\bar{\mu}} = \int_{\mathcal{Z}} m_{\mu_z} d\lambda(z)$. As a result, the optimal transport map is

$$T_{\mu_z \bar{\mu}} = T_{+m_{\bar{\mu}}} \circ T_{\mu'_z \bar{\mu}'} \circ T_{-m_{\mu_z}} \tag{32}$$

Remark 2.6 (Solution to (31)**)** *The non-linear matrix equation* (31) *has a unique solution that can be approached via the following iterative process:*

$$\int_{\mathcal{Z}} (\Sigma_i^{\frac{1}{2}} \Sigma_{\mu_z} \Sigma_i^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) \to \Sigma_{i+1}. \tag{33}$$

We refer interested readers to [4] for more details on the fixed point approach to the Wasserstein barycenter. The present work only applies this fact in the algorithm design in Section 6.

3. Wasserstein Barycenter Characterization of the Optimal Fair Learning Outcome

Optimal transport has been considered an adversarial or constrained optimization problem in its application to machine learning. In particular, some of the most popular unsupervised learning methods, such as K-means and PCA, are specific examples of the Wasserstein barycenter problems when putting restrictions on the admissible transport maps and relaxation on the weak equivalence requirement of the push-forwards w.r.t. test functions. See, for example, [39] for more details. But we apply optimal transport in an opposite direction so that the independence or imperceptibility of the sensitive variable Z becomes theoretically provable.

In this section, the primary goal is to develop the optimal affine map and pseudo-barycenter as tools to solve the challenge of the high computational cost of Wasserstein barycenter and optimal transport maps in high-dimensional data space. More specifically, we restrict the admissible transport maps to be merely affine maps while relaxing the fairness constraint to a sufficient and necessary level. The importance of efficiency in computing the barycenter and optimal transport maps will soon be clear in Section 4 when we compute the Pareto frontier along the Wasserstein geodesic path. Furthermore, the importance of affinity of transport maps will also be soon clear in Section 5 when solving the optimal fair data representation problem (3).

The organization of the current section is as follows: we first generalize the Wasserstein barycenter characterization of the optimal regression to all L^2-objective supervised learning models, then apply the optimal affine maps to estimate high-dimension optimal learning outcome. Now, we show that the (unique) solution to Problem 1 can be characterized as the Wasserstein barycenter of the conditional expectation sensitive marginals. The barycenter characterization of the optimal fair regression is first proved in [18, 24].

3.1 Wasserstein Barycenter Characterization

We start with a characterization of the optimal learning outcome of the L^2-objective supervised learning task. Let $\mathbb{E}(Y|X,Z)_z$ be the sensitive marginals of $(\mathbb{E}(Y|X,Z),Z)$ (or, equivalently, the sensitive conditionals of $\mathbb{E}(Y|X,Z)$ on $\{Z=z\}$ by Remark 2.2) for λ-a.e. $z \in \mathcal{Z}$, $\mathcal{L}(\mathbb{E}(Y|X,Z)_z) := \mu_z$, and $\bar{\mu}$ denote the Wasserstein barycenter of $\{\mu_z\}_{z\in\mathcal{Z}}$. Also, let $T(\cdot,z)$ denote the optimal transport map from μ_z to $\bar{\mu}$.

Lemma 3.1 (Optimal fair L^2-objective supervised learning characterization) *Assume that the conditional expectation marginals $\{\mu_z\}_{z\in\mathcal{Z}} \subset \mathcal{P}_{2,ac}(\mathcal{Y})$, then*

$$\overline{\mathbb{E}(Y|X,Z)} = T(\mathbb{E}(Y|X,Z),Z) := \{T(\mathbb{E}(Y|X,Z)_z,z)\}_{z\in\mathcal{Z}} \tag{34}$$

is the unique solution to Problem 1. Furthermore, we have

$$\begin{aligned}||Y - T(\mathbb{E}(Y|X,Z),Z)||_2^2 &= \inf_{f\in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})}\{||Y-f(X,Z)||_2^2 : f(X,Z)\perp Z\}\\ &= ||Y-\mathbb{E}(Y|X,Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\bar{\mu})d\lambda\end{aligned}$$

Proof First, notice that the fairness constraint $f(X,Z)\perp Z$ is equivalent to $\mathcal{L}(f(X,Z)_z) = \mu$ λ-a.e. for some $\mu \in \mathcal{P}(\mathcal{Y})$. Now, we prove the lower bound: let $f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})$ satisfies $f(X,Z)\perp Z$, we have

$$\begin{aligned}||Y-f(X,Z)||_2^2 =&||Y-\mathbb{E}(Y|X,Z)||_2^2 + ||\mathbb{E}(Y|X,Z)-f(X,Z)||_2^2\\ =&||Y-\mathbb{E}(Y|X,Z)||_2^2 + \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - f(X,Z)_z||_2^2 d\lambda\\ \geq&||Y-\mathbb{E}(Y|X,Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\mathcal{L}(f(X,Z)_z))d\lambda\\ \geq&||Y-\mathbb{E}(Y|X,Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\bar{\mu})d\lambda\end{aligned}$$

Here, the first line follows from the L^2 projection characterization of conditional expectation, the second follows from disintegration, the third from the definition of $\mathcal{W}_2$, and the fourth from the definition of the Wasserstein barycenter and the fairness restriction $f(X,Z) \perp Z$.

Next, we construct a $f_Y \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$ such that the lower bound is obtained. Let T_z denote the optimal transport map such that $(T_z)_\sharp \mu_z = \bar{\mu}$ for λ-a.e. $z \in \mathcal{Z}$. Define $T(\cdot, z) := T_z(\cdot)$ and

$$f_Y(X,Z) := T((\mathbb{E}(Y|X,Z),Z)). \tag{35}$$

Here, $\pi := (Id, T_z)_\sharp \mu_z d\lambda$ defines $\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Y} \times \mathcal{Y})$. Hence, we have $\pi = \pi_{(y,z)} d\lambda_{(\mathbb{E}(Y|X,Z),Z)}$ and $\pi_{(y,z)} = \delta_{T(y,z)}$ $\lambda_{(\mathbb{E}(Y|X,Z),Z)} - a.e.$. Since $(y,z) \to \pi_{(y,z)} = \delta_{T(y,z)}$ is $\mathcal{Y} \times \mathcal{Z}/\mathcal{P}(\mathcal{Y})$ measurable, we have $(y,z) \to T(y,z)$ is $\mathcal{Y} \times \mathcal{Z}/\mathcal{Y}$ measurable. It follows from $\mathbb{E}(Y|\cdot,\cdot) \otimes Id|_{\mathcal{Z}}(\cdot,\cdot) : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y} \times \mathcal{Z}$ being $\mathcal{X} \times \mathcal{Z}/\mathcal{Y} \times \mathcal{Z}$ measurable that $f_Y = T \circ (\mathbb{E}(Y|\cdot,\cdot) \otimes Id|_{\mathcal{Z}}(\cdot,\cdot))$ is $\mathcal{X} \times \mathcal{Z}/\mathcal{Y}$ measurable. Also, $\bar{\mu} \in \mathcal{P}_2(\mathcal{Y}) \implies ||f_Y(X,Z)||_2 < \infty$. This proves $f_Y \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$. It remains to show that the lower bound is obtained at $f_Y(X,Z)$. Indeed, by construction, we have

$$\begin{aligned}
||\mathbb{E}(Y|X,Z) - f_Y(X,Z)||_2^2 &= \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - f_Y(X,Z)_z||_2^2 d\lambda \\
&= \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T(\mathbb{E}(Y|X,Z)_z, z)||_2^2 d\lambda \\
&= \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, (T_z)_\sharp \mu_z) d\lambda \\
&= \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda.
\end{aligned}$$

It follows from the derivation of the lower bound above that

$$||Y - f_Y(X,Z)||_2^2 = \inf_{f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})} \{||Y - f(X,Z)||_2^2 : f(X,Z) \perp Z\} \tag{36}$$

Uniqueness follows from the uniqueness of $\bar{\mu}$ and the uniqueness of $T(\cdot, z)$. We are done. ■

The above result shows that the minimum L^2-loss for statistical parity can be nicely decomposed into two parts: (1) an $L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$ orthogonal projection loss due to the inference capability of (X,Z) w.r.t. Y and (2) an independence projection loss due to the statistical parity constraint. That is,

$$\underbrace{\inf_f \{||Y - f(X,Z)||_2^2 : f(X,Z) \perp Z\}}_{\text{minimum loss for statistical parity}} = \underbrace{||Y - \mathbb{E}(Y|X,Z)||_2^2}_{\text{orthogonal projection loss}} + \underbrace{\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda}_{\text{independence projection loss}} .$$

Furthermore, to construct the optimal fair L^2 learning outcome, one first performs L^2 orthogonal projection to obtain the conditional expectation $\mathbb{E}(Y|X,Z)$, then outputs the Wasserstein barycenter of the sensitive marginals of $\mathbb{E}(Y|X,Z)$ as the optimal (with respect to L^2-objective) fair (for statistical parity) result.

Unfortunately, in practice, the characterization suffers from a lack of efficient methods to compute the Wasserstein barycenter and obtain an explicit formula of the optimal transport maps [3]. Current methods restrict the sensitive variable Z to be binary mainly because the computation of a multi-marginal barycenter is expensive. Furthermore, notice the current methods restrict the dependent variable Y to be one-dimensional, because the only well-known exact solution to transport maps is the inverse of cumulative function that merely works for one-dimensional variables.

Therefore, to provide methods using the characterization in high-dimensional dependent variable cases, we introduce the optimal affine map and the associated pseudo-barycenter.

3.2 Optimal Affine Estimation: Pseudo-barycenter

To solve the challenge of deriving an explicit formula for the Wasserstein barycenter and optimal transport maps, we restrict the admissible transport maps to be affine and show that the estimation of the Wasserstein barycenter via optimal affine maps coincides with the true Wasserstein barycenter in the Gaussian case, and that the estimation error is bounded in the case of general distributions. In other words, we consider the choice of positive definite affine maps under two circumstances:

1 We assume the marginals are non-degenerate Gaussian. That is, $\{\mathbb{E}(Y|X,Z)_z\}_z$ are assumed to be non-degenerate Gaussian vectors λ-a.e..

2 Instead of making assumptions on the data distribution, we relax the independence constraint to the independence between Z and merely the first two moments of $f(X,Z)$.

From a theoretical perspective, affine maps allow us to derive (nearly) closed-form solutions under either of the assumptions mentioned above. Also, affine maps allow us to develop a pre-processing approach by directly applying the obtained maps to the original data before training, even though such maps are constructed to push the post-training marginals toward their barycenter.

From a practical perspective, the advantage is obvious: the computation of affine maps only uses (sample estimation of) the first two moments of the marginal distributions and hence is highly efficient compared to the computation of general Brenier's maps, especially in the case of high-dimension data.

Before developing the pseudo-barycenter, the following remarks compare in more detail the exact barycenter with its affine approximation.

Remark 3.1 (Applying pseudo-barycenter vs exact barycenter) *The comparison between the pseudo-barycenter method and the exact barycenter is an analog of the comparison between the linear regression model and the exact conditional expectation: When there is no worry about over-fitting, a practitioner who cares more about the strict goal of minimizing L^2 error (analog: the strict statistical parity guarantee) should always try to find the exact conditional expectation function (analog: the exact barycenter and the corresponding exact transport maps) by using more complicated models. But the simplicity, robustness, and interpretability of linear regression (analog: pseudo-barycenter and optimal affine maps) are often useful in practice.*

We define the pseudo-barycenter, using merely matrix calculations, as follows:

Definition 3.1 *The post-processing pseudo-barycenter $\hat{Y}^{\dagger}$ is given via*

$$\hat{Y}^{\dagger} := T_{\mathit{affine}}(\hat{Y}, Z), \tag{37}$$

where

$$T_{\mathit{affine}}(\cdot, z) := \Sigma_{\hat{Y}_z}^{-\frac{1}{2}}(\Sigma_{\hat{Y}_z}^{\frac{1}{2}}\Sigma\Sigma_{\hat{Y}_z}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{\hat{Y}_z}^{-\frac{1}{2}}, \tag{38}$$

and Σ is the unique solution to

$$\int_{\mathcal{Z}}(\Sigma^{\frac{1}{2}}\Sigma_{\hat{Y}_z}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}d\lambda(z) = \Sigma. \tag{39}$$

To obtain (an approximation of) the unique solution, we apply the iterative method (33) in Remark 2.6 when designing our algorithm in Section 6.

Now, Lemma 2.2 shows that under the assumption of Gaussianity of the learning outcome marginals, the optimal transport map is affine and the pseudo-barycenter is indeed the Wasserstein barycenter. Moreover, Lemma 2.3 shows that the barycenter of Gaussian marginals is still Gaussian. Therefore, the optimal maps from the marginals to the barycenter are determined entirely by the first two moments.

Lemma 3.2 (Post-processing pseudo-barycenter in the Gaussian case) *Assume $\hat{Y}_z \sim \mathcal{N}(0, \Sigma_z)$ for λ-a.e. $z \in \mathcal{Z}$, then $\hat{Y}^{\dagger}$ is the Wasserstein barycenter of $\{\hat{Y}_z\}_z$.*

It follows from Theorem 3.2 that, if $\hat{Y} = \mathbb{E}(Y|X, Z)$, then $Y^{\dagger}$ is the solution to the Wasserstein barycenter characterization of the optimal fair learning outcome.

Finally, we show that the pseudo-barycenter is the optimal affine estimation of the Wasserstein barycenter in the case of general marginal distributions. To do so, we need to first put restrictions on the admissible transport maps. However, such a restriction on admissible maps leads to a necessary relaxation of the fairness constraint. To see the necessity, Lemma 2.2 shows positive definite affine maps transform distributions within the same location-scale family. Therefore, given marginals Y_1 and Y_2 from different location-scale families, affine maps are not able to transform them to each other. That implies the non-existence of the barycenter under the original independence restriction. Indeed, if a barycenter of $\{Y_z\}_{z\in\{1,2\}}$ exists under the restriction of positive definite affine maps, then Y_1 and Y_2 belong to the same location-scale family as their barycenter, which contradicts the assumption of general distributions. That is, the Wasserstein barycenter characterization does not have a solution when we admit merely affine transport maps in the general marginal distribution case.

On the other hand, notice that the best affine maps can achieve is to map Y_1 to a Y_2', which shares the same first two moments with Y_2 within the Y_1 location-scale family. We call such Y_2' a Y_1 location-scale family analog of Y_2. Therefore, we propose the following relaxation of the fairness constraint that suffices to guarantee the existence of a solution to the relaxed version of (1) with merely positive definite affine transport maps:

$$m_{f(X,Z)}, \Sigma_{f(X,Z)} \perp Z \tag{40}$$

where $m_{f(X,Z)}$, and $\Sigma_{f(X,Z)}$, denotes respectively the first and second moment of $f(X, Z)$.

Remark 3.2 (Fairness guarantee of the relaxation) *The adversarial task of testing and exploiting probabilistic independence between $f(X,Z)$ and Z is equivalently difficult to enforcing the independence. One common strategy is to explore its equivalence to the independence between all moments of $f(X,Z)$ and Z, provided the boundedness of the two random variables. But the verification or enforcement of independence among higher moments is extremely vulnerable to data noise in practice. Thus, instead of enforcing $f(X,Z) \perp Z$, one could relax the constraint to the independence between Z and some of the moments of $f(X,Z)$. In this section, we focus on the first two moments. That is, $m_{f(X,Z)}, \Sigma_{f(X,Z)}$ where $m_{f(X,Z)} := \mathbb{E}(f(X,Z))$ and $\Sigma_{f(X,Z)} := \mathbb{E}((f(X,Z)-\mathbb{E}(f(X,Z)))(f(X,Z)-\mathbb{E}(f(X,Z)))^T)$. It is not hard to notice that the relaxation is already strong enough to result in imperceptibility to any unsupervised learning algorithm that uses merely the mean and covariance of data to extract information, such as K-means and PCA.*

Therefore, the optimal affine estimation of the Wasserstein barycenter characterization is given by:

Problem 4 (Optimal affine estimation of barycenter problem)

$$\inf_{f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})} \{||Y - f(X,Z)||_2^2 : m_{f(X,Z)}, \Sigma_{f(X,Z)} \perp Z\}. \tag{41}$$

Now, we show that the pseudo-barycenter defined above is indeed the solution to Problem 4 and hence the optimal affine estimate of the optimal fair learning outcome. To prove the main result, we need the following result: given any fixed covariance matrix, the optimal positive definite affine maps result in the lowest Wasserstein distance such that the push-forwards all share the same fixed covariance matrix. To simplify notation, let $\mu_z := \mathcal{L}(\mathbb{E}(Y|X,Z)_z)$. Also, let $m_{Y|X_z}$ and $\Sigma_{Y|X_z}$ denote the mean and covariance matrix of $\mathbb{E}(Y|X,Z)_z$ respectively.

Lemma 3.3 (Projection Lemma) *Assume $\{\mu_z\}_z \subset \mathcal{P}_{2,ac}(\mathcal{Y})$. If $m_{Y|X_z} = 0, \Sigma_{Y|X_z} \succ 0$ λ-a.e., for any $\Sigma \succ 0$,*

$$\inf_{\hat{Y}:\Sigma_{\hat{Y}_z}=\Sigma} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mathcal{L}(\hat{Y}_z))d\lambda(z) \tag{42}$$

admits a unique solution, denoted by $\hat{Y}_\Sigma$, that satisfies

$$\hat{Y}_{\Sigma,z} := T_\Sigma(\hat{Y}_z, z) \tag{43}$$

where $T_\Sigma(\cdot,z) := \Sigma_{Y|X_z}^{-\frac{1}{2}}(\Sigma_{Y|X_z}^{\frac{1}{2}}\Sigma\Sigma_{Y|X_z}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{Y|X_z}^{-\frac{1}{2}}$.

Proof

$$\begin{aligned}\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mathcal{L}(\hat{Y}_z)d\lambda(z) &= \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T_\Sigma(\hat{Y}_z, z)||_2^2 d\lambda(z)\\ &= \int_{\mathcal{Z}} \inf_{\nu:\Sigma_\nu=\Sigma} \mathcal{W}_2^2(\mu_z,\nu)d\lambda(z)\\ &= \inf_{\nu:\Sigma_{\nu_z}=\Sigma} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z,\nu_z)d\lambda(z),\end{aligned}$$

where the second equality follows from the characterization of Gelbrich's bound, see for example Proposition 2.4 in [20]. Now, let $\hat{Y}' \neq \hat{Y}_\Sigma$ but also satisfy $\Sigma_{\hat{Y}'} = \Sigma$ λ-a.e., then we have

$$\begin{aligned}\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \hat{Y}_{\Sigma,z}||_2^2 d\lambda(z) &< \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mathcal{L}(\hat{Y}_z')) d\lambda(z) \\ &\leq \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \hat{Y}_z'||_2^2 d\lambda(z),\end{aligned}$$

where the first inequality is strict due to the uniqueness of Brenier's maps $T_\Sigma(\cdot, z)$ and hence of $T_\Sigma(\hat{Y}_z, z)$ λ-a.e.. The proof is complete. ■

Remark 3.3 (Intuition of the Projection Lemma) *Intuitively, for an arbitrary positive definite matrix Σ, one can consider $T_\Sigma(\cdot, z)$ as the projection map (w.r.t. $\mathcal{W}_2$ distance) onto*

$$\{\nu \in \mathcal{P}_2(\mathcal{Y}) : \Sigma_\nu = \Sigma\} \tag{44}$$

which is the set of centered probability measures with fixed covariance matrix Σ in $(\mathcal{P}_2(\mathcal{Y}), \mathcal{W}_2)$. In other words, given a probability measure, the maps $\{T_\Sigma(\cdot, z)\}_z$ finds the closest (w.r.t. the Wasserstein distance) point in the set for each of the marginals.

Finally, we are ready to prove the justification of the pseudo-barycenter in the case of general distributions.

Theorem 3.1 (Optimal affine estimation of $\mathcal{W}_2$ barycenter: Pseudo-barycenter) $\mathbb{E}(Y|X,Z)^\dagger := \{T_{affine}(\mathbb{E}(Y|X,Z)_z, z)\}_z$ *is the unique solution to Problem 4:*

$$\inf_{f \in L^2(\mathcal{X}\times\mathcal{Z},\mathcal{Y})} \{||Y - f(X,Z)||_2^2 : m_{f(X,Z)}, \Sigma_{f(X,Z)} \perp Z\}, \tag{45}$$

provided $\{\mu_z\}_z \subset \mathcal{P}_{2,ac}(\mathcal{Y})$.

Proof First, we fix $\Sigma \succ 0$ arbitrary and denote $\hat{Y}_{\Sigma,z} := T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)$ for λ-a.e. $z \in \mathcal{Z}$, we have

$$||Y - T_\Sigma(\mathbb{E}(Y|X,Z), Z)||_2^2 - ||Y - \mathbb{E}(Y|X,Z)||_2^2 = \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \hat{Y}_{\Sigma,z}||_2^2 d\lambda(z) \tag{46}$$

and it follows from Lemma 3.3 that

$$\begin{aligned}\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \hat{Y}_{\Sigma,z}||_2^2 d\lambda(z) &= \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mathcal{L}(T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)) d\lambda(z) \\ &= \min_{\nu:\Sigma_{\nu_z}=\Sigma} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \nu_z) d\lambda(z).\end{aligned}$$

Therefore, (41) boils down to the following:

$$\inf_{\Sigma \succ 0} \Big\{ \int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)||_2^2 d\lambda(z) \Big\}. \tag{47}$$

Finally, notice that

$$
\begin{aligned}
&\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)||_2^2 d\lambda(z) \\
=&\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z||_2^2 + ||T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)||_2^2 - 2\langle \mathbb{E}(Y|X,Z)_z, T_\Sigma(\mathbb{E}(Y|X,Z)_z, z)\rangle_2 d\lambda(z) \\
=&\int_{\mathcal{Z}} \text{Trace}(\Sigma_{Y|X_z}) + \text{Trace}(\Sigma) - 2\mathbb{E}(\mathbb{E}(Y|X,Z)_z^T T_\Sigma(\mathbb{E}(Y|X,Z)_z, z) d\lambda(z) \\
=&\int_{\mathcal{Z}} \text{Trace}(\Sigma_{Y|X_z}) + \text{Trace}(\Sigma) - 2\langle T_\Sigma, \Sigma_{Y|X_z}\rangle_F d\lambda(z) \\
=&\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)'_z - T_\Sigma(\mathbb{E}(Y|X,Z)'_z, z)||_2^2 d\lambda(z),
\end{aligned}
$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product and $X' \sim \mathcal{N}(m_X, \Sigma_X)$ denotes the Gaussian analog of X. It follows from definition of $T_{\text{affine}}(\mathbb{E}(Y|X,Z)_z, z)$ with $T_{\text{affine}}(\cdot, z) := \Sigma_{Y|X_z}^{-\frac{1}{2}}(\Sigma_{Y|X_z}^{\frac{1}{2}}\Sigma\Sigma_{Y|X_z}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{Y|X_z}^{-\frac{1}{2}}$ and Lemma 2.3 that $\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \mathbb{E}(Y|X,Z)_z^\dagger||_2^2 d\lambda(z)$ is the unique lower bound of the objective function in (47). It then follows from the uniqueness of Brenier's map that $\mathbb{E}(Y|X,Z)^\dagger$ is the unique solution to (41). ■

In this section, we focus on applying the optimal affine transport map and the pseudo-barycenter to find a computationally efficient estimation of the optimal fair learning outcome in high-dimensional space. As we mentioned above, it will soon become clear in the next two sections and numerical experiments that a combination of McCann interpolation and optimal affine maps in matrix form results in not only a mathematically neat solution to estimate the Pareto frontier, which significantly reduces computational expense in practice, but also a necessary tool to help us circumvent the post-processing nature and solve the optimal fair data representation problem (3).

Now, we are ready to address the lack of a precise theoretical characterization of the Pareto frontier between utility and fairness, which turns out to be a natural generalization of the Wasserstein barycenter characterization of the optimal fair L^2-objective learning outcome.

4. Wasserstein Geodesics Characterization of Pareto Frontier

In reality, rather than looking for the optimal fair learning outcome, practitioners may have to choose a middle ground: sacrificing some prediction accuracy while tolerating a certain level of disparity. Therefore, it is tempting to generalize the barycenter characterization of the optimal fair learning outcome to the entire Pareto frontier between prediction error and statistical disparity. In this section, we show that the constant-speed geodesics from the conditional expectation sensitive marginals to their Wasserstein barycenter characterize the Pareto frontier on the Wasserstein space, in which utility loss and statistical disparity are quantified respectively by the L^2 norm and the average pair-wise Wasserstein distance among the sensitive marginals. As a result, given the optimal transport maps, one can derive a closed-form solution to the geodesics and thereby the Pareto frontier using McCann interpolation.

Here, we first provide a post-processing characterization of the Pareto frontier, Theorem 4.1, which is of theoretical interest and great generality. Then, we derive a closed-form solution to Problem 2 based on this characterization. The results form a direct generalization of the barycenter characterization, which is Lemma 3.1, and practitioners can apply the result together with the pseudo-barycenter and McCann interpolation to obtain the optimal affine estimation to the post-processing Pareto frontier. Later in Section 5, we further apply the result to provide a characterization of the exact solution and an optimal affine estimation of the solution to the optimal fair data representation problem (3).

Now, we start to characterize the Pareto frontier. In the rest of the section, we denote $\mathcal{L}(\mathbb{E}(Y|X,Z)) =: \mu, \mathcal{L}(\mathbb{E}(Y|X,Z)_z) =: \mu_z$. For utility, given any measurable function $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, we define the increased prediction error by the L^2-norm of the difference between $f(X,Z)$ and the orthogonal projection $\mathbb{E}(Y|X,Z)$:

$$L(f(X,Z)) := ||\mathbb{E}(Y|X,Z) - f(X,Z)||_2 = (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - f(X,Z)_z)||_2^2 d\lambda(z))^{\frac{1}{2}}. \quad (48)$$

To simplify notation, we also denote

$$L(T') := L(T'(\mathbb{E}(Y|X,Z),Z)) = (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T'_z(\mathbb{E}(Y|X,Z)_z)||_2^2 d\lambda(z))^{\frac{1}{2}}. \quad (49)$$

for any measurable $T' : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$.

To relax the hard independence constraint for the Pareto frontier, we quantify the statistical disparity of a given learning outcome or prediction $\hat{Y}$ by the average pairwise Wasserstein distance among its sensitive marginals:

Definition 4.1 (Wasserstein disparity)

$$D(\hat{Y},Z) := \Big(\int_{\mathcal{Z}^2} \mathcal{W}_2^2(\mathcal{L}(\hat{Y}_{z_1}), \mathcal{L}(\hat{Y}_{z_2})) d\lambda(z_1) d\lambda(z_2)\Big)^{\frac{1}{2}}. \quad (50)$$

In our setting, $\hat{Y} = f(X,Z)$ for some $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$. Also, to simplify notation, we denote the Wasserstein disparity that remains in the already deformed (by applying T') conditional expectation by

$$D(T') := D(T'(\mathbb{E}(Y|X,Z),Z),Z) = (\int_{\mathcal{Z}^2} \mathcal{W}_2^2((T'_{z_1})_\sharp \mu_{z_1}, (T'_{z_2})_\sharp \mu_{z_2}) d\lambda(z_1) d\lambda(z_2))^{\frac{1}{2}} \quad (51)$$

for any measurable $T' : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$. Here, $T_z = T(\cdot, z) : \mathcal{Y} \to \mathcal{Y}$ for λ-a.e. $z \in \mathcal{Z}$.

We adopt the Wasserstein disparity as a statistical disparity quantification due to the following desirable properties:

- **Wasserstein disparity characterizes statistical parity:**

$$\mathcal{D}(f(X,Z),Z) = 0 \iff f(X,Z) \perp Z$$

- **Physics interpretation:** Due to the definition based on the Wasserstein distance, Wasserstein disparity can be understood as the expected minimum amount of work that is required to move one randomly chosen marginal to another random chosen one. Therefore, the larger $\mathcal{D}(f(X,Z),Z)$ is, the more necessary work is expected to remove the distributional discrepancy among the sensitive groups on $f(X,Z)$.

Now, let $T : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$ satisfy $T(\cdot, z)$ being the optimal transport maps from $\{\mu_z\}_z$ to their barycenter $\bar{\mu}$ for λ-a.e. $z \in \mathcal{Z}$ (See construction of T in the proof of Lemma 3.1), we define

$$V := L(T) = (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T(\mathbb{E}(Y|X,Z)_z, z)||_2^2 d\lambda(z))^{\frac{1}{2}} \tag{52}$$

$$= (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \overline{\mathbb{E}(Y|X,Z)}_z||_2^2 d\lambda(z))^{\frac{1}{2}}. \tag{53}$$

As shown in Lemma 3.1, V is the minimum increase of L^2 error (or, in physics, the minimum work/energy required) to deform $\mathbb{E}(Y|X,Z)$ to satisfy statistical parity. Before showing the main result, we need to define the geodesic on metric space to show the explicit form of constant speed geodesic on the Wasserstein space, which plays a key role in the proof.

Definition 4.2 (Constant-speed geodesic between two points on metric space) *Given a metric space (X, d) and $x, x' \in X$, the constant-speed geodesic between x and x' is a continuously parametrized path $\{x_t\}_{t\in[0,1]}$ such that $x_0 = x$, $x_1 = x'$, and $d(x_s, x_t) = |t-s|d(x,x'), \forall s,t \in [0,1]$.*

The following lemma, which is well known as the McCann (displacement) interpolation [41, Chapter 7] in the optimal transport literature, shows that a linear interpolation using the optimal transport plan results in the constant-speed geodesic on the Wasserstein space.

Lemma 4.1 (Constant-speed geodesic on Wasserstein space, [32, 41]) *Given $\mu_0, \mu_1 \in (\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ and γ the optimal transport plan in between, let $\pi_t(x,y) := (1-t)x + ty$, then*

$$\mu_t := (\pi_t)_\sharp \gamma, t \in [0,1] \tag{54}$$

is the constant-speed geodesic between μ_0 and μ_1.

Proof See Appendix B ■

Remark 4.1 (Linear interpolation formula for $\mathcal{W}_2$ deodesics) *If there exists an optimal transport map T such that $T_\sharp(\mu_0) = \mu_1$, then the McCann interpolation has the simple form*

$$\mu_t = ((1-t)Id + tT)_\sharp \mu_0, t \in [0,1]. \tag{55}$$

We apply this simple formula to obtain a closed-form estimation of the Pareto frontier in algorithm design, see Section 6.

Now, we are ready to establish the main result, which shows that V is a lower bound of $L(f(X,Z)) + \frac{1}{\sqrt{2}}D(f(X,Z),Z)$ for any measurable function $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and is achieved along the constant-speed geodesics from the sensitive marginals of the conditional expectation to their barycenter on the Wasserstein space.

Theorem 4.1 **($\mathcal{W}_2$ geodesics characterization of a linear Pareto frontier)** *Define L, D, V as above and assume $\mu_z \in \mathcal{P}_{2,ac}(\mathcal{Y}), \lambda - a.e.$. It follows that*

$$V \leq L(f(X,Z)) + \frac{1}{\sqrt{2}} D(f(X,Z),Z) \tag{56}$$

for any measurable function $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$. Furthermore, define $T(t)$ such that $T(t)(\cdot, z) := (1-t)Id + t(T(\cdot, z)), t \in [0,1]$ is the linear interpolation between the identity map and the optimal transport map for λ-a.e. $z \in \mathcal{Z}$, then equality holds in (56) if and only if $f(X,Z) = T(t)(\mathbb{E}(Y|X,Z),Z), t \in [0,1]$ as

$$L(T(t)) = tL(T(1)) = tV \tag{57}$$

$$\frac{1}{\sqrt{2}} D(T(t)) = \frac{1}{\sqrt{2}}(1-t)D(T(0)) = (1-t)V. \tag{58}$$

Proof See Appendix B. ■

Remark 4.2 (Intuition of Theorem 4.1: a Euclidean analog) *Here, we provide a Euclidean analog of Theorem 4.1. In fact, our proof is based on the observation of the analog and equivalent to it when one considers $x \to \delta_x$ as an embedding from $\mathcal{X}$ to $\mathcal{P}_2(\mathcal{X})$.*

Let $X := \{x_i\}_{i=1}^N$ be a fixed data set on the Euclidean space $\mathcal{X}$ ($N = 3$ in Figure 2), $\tilde{X} := \{\tilde{x}_i\}_{i=1}^N$ be a data set consisting of N arbitrarily chosen data points on $\mathcal{X}$, and define the following:

- *[Euclidean analog of V] std$(X) := (\frac{1}{N}\sum_{i=1}^N ||x_i - m_x||^2)^{\frac{1}{2}}$ with $m_x := \frac{1}{N}\sum_{i=1}^N x_i$,*
- *[Euclidean analog of L] solid$(\tilde{X}) := (\frac{1}{N}\sum_{i=1}^N ||x_i - \tilde{x}_i||^2)^{\frac{1}{2}}$,*
- *[Euclidean analog of D] dotted$(\tilde{X}) := (\frac{1}{N^2}\sum_{i,j=1}^N ||\tilde{x}_i - \tilde{x}_j||^2)^{\frac{1}{2}}$.*

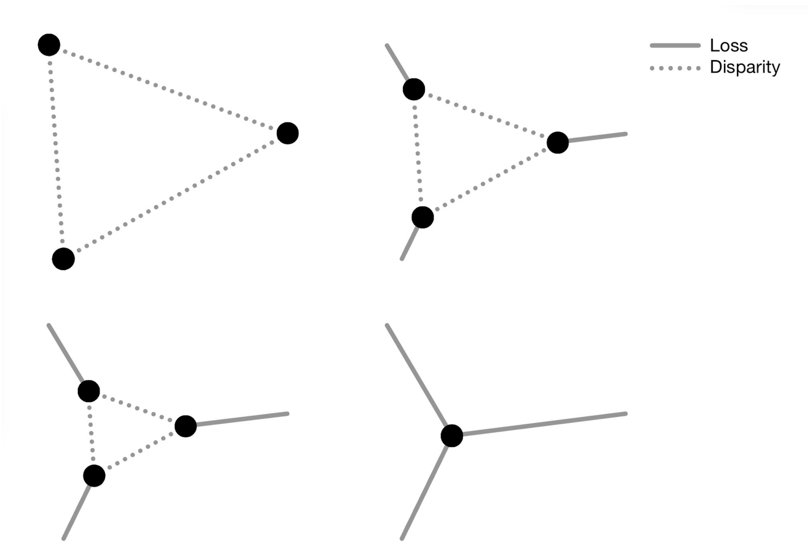


Figure 2: In this figure, we have three data points on an Euclidean space traveling along straight lines (Euclidean geodesics) to their average (Euclidean barycenter). Define (1) std := the standard deviation of the three points, (2) solid line (loss) := the average moving (Euclidean) distance away from their original location, and (3) dotted line (disparity) := the average pairwise (Euclidean) distance among them. One can show that std $\leq$ solid + $\frac{1}{\sqrt{2}}$dotted where equality holds if and only if the three points travel at constant-speed along straight lines to their average.

It is straight-forward to verify that (1)

$$std(X) \leq \underbrace{solid(\tilde{X})}_{utility\ loss} + \frac{1}{\sqrt{2}} \underbrace{dotted(\tilde{X})}_{disparity},$$

and (2) equality holds if and only if $\tilde{X} = X(t) := \{(1-t)x_i + tm_x\}_{i=1}^N$ *for* $t \in [0,1]$ *as* $loss(X(t)) = tstd(X)$ *and* $\frac{1}{\sqrt{2}}disparity(X(t)) = (1-t)std(X)$.

Since V (the minimum work or energy required for statistical parity) is fixed for the data (X, Y, Z) when one applies (X, Z) to predict Y, the above theorem implies that the Pareto frontier between the increased prediction error $L(T)$ and the remaining disparity $D(T)$ is a line that results from the constant-speed geodesics from the marginal conditional expectations to their barycenter on the Wasserstein space. In particular, let $T(t)(\mathbb{E}(Y|X,Z), Z) := \{T(t)(\mathbb{E}(Y|X_z), z)\}_z$, λ-a.e., $t \in [0,1]$, we arrive at a closed-form solution to Problem 2:

Corollary 4.1 (Pareto optimal fair L^2-objective learning) *Given* (X, Y, Z) *satisfying* $\mu_z \in \mathcal{P}_{ac}$, λ*-a.e., then*

$$f_d(X,Z) := \begin{cases} T(1-\frac{d}{\sqrt{2}V})(\mathbb{E}(Y|X,Z), Z), & \text{if } d \in [0, \sqrt{2}V] \\ \mathbb{E}(Y|X,Z), & \text{if } d \in (\sqrt{2}V, \infty) \end{cases} \tag{59}$$

are the unique solutions to Problem 2 for $d \in [0, \infty)$.

Proof If $d \in (\sqrt{2}V, \infty)$, then it follows from Theorem 4.1 that $D(\mathbb{E}(Y|X,Z)) = D(T(0)) = \sqrt{2}V < d$. Hence, Problem 2 reduces to the unconstrained L^2 projection problem and the optimal solution is $\mathbb{E}(Y|X,Z)$. Now, for a fixed $d \in [0, \sqrt{2}V]$, assume for contradiction that $\exists f \in L^2(\mathcal{X} \times \mathcal{Z}, \mathcal{Y})$ such that

$$||Y - f(X,Z)||_2^2 < ||Y - T(t)(\mathbb{E}(Y|X,Z), Z)||_2^2$$

for $t = 1 - \frac{d}{\sqrt{2}V}$. Then, let $\overline{f(X,Z)}$ denote the Wasserstein barycenter of $\{f(X,Z)_z\}_z$, we have

$$\begin{aligned}
||Y - \overline{f(X,Z)}||_2^2 &\leq ||Y - f(X,Z)||_2^2 + ||f(X,Z) - \overline{f(X,Z)}||_2^2 \\
&< ||Y - T(t)(\mathbb{E}(Y|X,Z), Z)||_2^2 + ||f(X,Z) - \overline{f(X,Z)}||_2^2 \\
&= ||Y - \mathbb{E}(Y|X,Z)||_2^2 + L(T(t)) + \frac{1}{\sqrt{2}}D(f(X,Z)) \\
&= ||Y - \mathbb{E}(Y|X,Z)||_2^2 + (V - \frac{1}{\sqrt{2}}d) + \frac{1}{\sqrt{2}}d \\
&= ||Y - \mathbb{E}(Y|X,Z)||_2^2 + V
\end{aligned}$$

where the second line follows from the assumption, the third from L^2 orthogonal decomposition and Theorem 4.1, and the forth from the assumption and Theorem 4.1. The strict inequality above contradicts the optimality of $\overline{\mathbb{E}(Y|X,Z)}$ shown in Lemma 3.1. That proves the optimality of $T(1 - \frac{d}{\sqrt{2}V})(\mathbb{E}(Y|X,Z), Z)$ for the fixed d. Uniqueness result follows from the uniqueness of $\overline{\mathbb{E}(Y|X,Z)}$ shown in Lemma 3.1. Since the choice of $d \in [0, \sqrt{2}V]$ is arbitrary, we are done. ■

We note that Corollary 4.1 together with Lemma 4.1 and Remark 4.1 provide a post-processing approach to (estimate) the Pareto frontier: applying McCann interpolation to the Brenier's maps between the learning outcome sensitive marginals $\{\mathbb{E}(Y|X,Z)_z\}_z$ and their (pseudo-) barycenter. One can apply Algorithm 1 directly with the learning outcome marginals as inputs.

From a theoretical perspective, various metrics of disparity that differ from D, the Wasserstein disparity (Definition 4.1), can be used and the theoretical results derived in this section provide a lower bound estimation for the Pareto frontier that uses other disparity metrics. The quality of the lower bound can be studied using the relationship between the Wasserstein distance and the defined disparity metric. Also, the present work provides a numerical study on the lower bound estimation in Section 6 to which we refer the interested readers for more details.

In practice, various metrics of disparity are adopted, such as the prediction success ratio (difference from 1) in classification [13] and the Kolmogorov-Smirnov distance for 1-dimensional regression [18]. The proposed estimation of the Pareto frontier leaves the choice of α to practitioners who would face specific fairness requirements and disparity metrics.

5. Optimal Fair Data Representation for Supervised Learning

In this section, we study the optimal fair data representation problem, Problem 3, that is motivated by the current challenges in the pre-processing or synthetic data design approach to fair machine learning. To solve the problem, we first characterize the exact solution using a dependent and independent Wasserstein barycenter pair, see Lemma 5.3. Then, we define a dependent and independent pseudo-barycenter pair via optimal affine maps, and prove that the pair is the exact optimal fair data representation with Gaussian marginals, cf. Lemma 5.5 and the optimal affine estimate of the representation with general marginals in Theorem 5.2.

5.1 Wasserstein Barycenter Pair Characterization

We will prove a characterization of the solutions to Problem 3. To start, notice that since $(\tilde{X}, Z) = T \otimes Id|_{\mathcal{Z}}(X, Z)$ for some measurable map $T \otimes Id|_{\mathcal{Z}} : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$, we have $\sigma((\tilde{X}, Z)) \subset \sigma((X, Z))$. Also, from $\tilde{X} \perp Z$, we have $\sigma(\tilde{X}) \subset \sigma(\tilde{X}) \otimes \sigma(Z) = \sigma((\tilde{X}, Z))$. Therefore, $\sigma(\tilde{X}) \subset \sigma((X, Z))$ and it follows from L^2 orthogonal decomposition that

$$||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 = ||Y - \mathbb{E}(Y|X, Z)||_2^2 + ||\mathbb{E}(Y|X, Z) - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2. \tag{60}$$

The first term on the right hand side can be interpreted as the minimum loss of information by using (X, Z) to predict Y. Furthermore, one can decompose the second term on the right hand side of (60):

$$\begin{aligned} &||\mathbb{E}(Y|X, Z) - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 \\ =&||\mathbb{E}(Y|X, Z) - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 + ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 \\ =&||\mathbb{E}(Y|X, Z) - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 + \int_{\mathcal{Z}} ||\mathbb{E}(Y_z|\tilde{X}) - \mathbb{E}(\tilde{Y}|\tilde{X})_z||_2^2 d\lambda(z). \end{aligned}$$

Here, the first equality follows from L^2 orthogonal decomposition. The second equality follows from disintegration, the fairness constraint $\tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X}) \perp Z$, and the fact that $\tilde{X} \perp Z$ implies

$$\mathbb{E}(Y_z|\tilde{X}) = \mathbb{E}(Y|\tilde{X}, Z)_z.$$

See Appendix C for the proof.

Now, the key observation is that, given a fixed $\tilde{X} \perp Z$, the choice of $\tilde{Y}$ depends only on the second term on the right, which forms a Wasserstein barycenter problem with marginals being $\{\mathbb{E}(Y_z|\tilde{X})\}_z$. Hence, the optimal choice of $\tilde{Y}$ is the one which satisfies $\mathbb{E}(\tilde{Y}|\tilde{X}) = \overline{\mathbb{E}(Y|\tilde{X}, Z)}$, where $\overline{\mathbb{E}(Y|\tilde{X}, Z)})$ is the Wasserstein barycenter of $\{\mathbb{E}(Y_z|\tilde{X})\}_z$. Therefore, we denote the optimal choice of $\tilde{Y}$ to be $\bar{Y}$ which satisfies $\mathbb{E}(\bar{Y}|\tilde{X}) = \overline{\mathbb{E}(Y|\tilde{X}, Z)}$.

It remains to find the optimal choice of $\tilde{X}$. The following result shows that the optimal choice is the one admissible $\tilde{X}$ which generates the finest sigma-algebra.

Lemma 5.1 (Finer sigma-algebra, more accurate optimal fair learning) *Let $\tilde{X}, \tilde{X}' \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : \tilde{X} \perp Z\}$. If $\sigma(\tilde{X}') \subset \sigma(\tilde{X})$, then*

$$||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}|\tilde{X})||_2^2 \leq ||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}'|\tilde{X}')||_2^2 \tag{61}$$

where $\bar{Y}$ and $\bar{Y}'$ satisfy $\mathbb{E}(\bar{Y}|\tilde{X}) = \overline{\mathbb{E}(Y|\tilde{X}, Z)}$ and $\mathbb{E}(\bar{Y}'|\tilde{X}') = \overline{\mathbb{E}(Y'|\tilde{X}', Z)}$.

Proof See Appendix C. ■

Therefore, it is clear that our optimal choice of $\tilde{X}$ is the one that generates the finest sigma-algebra while satisfying $\tilde{X} \perp Z$. The following technical lemma shows that the barycenter of $\{X_z\}_{z\in\mathcal{Z}}$ is one of the optimal choices.

Lemma 5.2 ($\bar{X}$ generates the finest sigma-algebra among admissible) *If $\{\mathcal{L}(X_z)\}_z \subset \mathcal{P}_{2,ac}(\mathcal{X})$ λ-a.e., then $\sigma((\bar{X}, Z)) = \sigma((X, Z))$. In addition, $\sigma(\tilde{X}) \subset \sigma(\bar{X})$ for all $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : \tilde{X} \perp Z\}$.*

Proof See Appendix C. ■

Therefore, Lemma 5.1, Lemma 5.2, and the choice of $\bar{Y}$ above together provide a characterization of the solution to Problem 3.

Lemma 5.3 (Characterization of optimal fair data representation) *Let $\bar{X}$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ denote the respective Wasserstein barycenter of $\{X_z\}_z$ and $\{\mathbb{E}(Y_z|\bar{X})\}_z$. If $\{\mathcal{L}(X_z)\}_z \subset \mathcal{P}_{2,ac}(\mathcal{X})$ and $\{\mathcal{L}(\mathbb{E}(Y|\bar{X}, Z)_z)\}_z \subset \mathcal{P}_{2,ac}(\mathcal{Y})$, then the following are equivalent:*

- $(\tilde{X}, \tilde{Y}) \in \arg\min_{(\tilde{X},\tilde{Y})\in\mathcal{D}}\{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : \tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X}, Z) \perp Z\}$.
- $(\tilde{X}, \tilde{Y}) \in \{(\tilde{X}, \tilde{Y}) \in \mathcal{D} : \sigma(\tilde{X}) = \sigma(\bar{X}), \mathbb{E}(\tilde{Y}|\bar{X}) = \overline{\mathbb{E}(Y|\bar{X}, Z)}\}$.

In Lemma 5.3, the choice of $\bar{X}$ is not unique. In fact, any random variable $\tilde{X}$ that satisfies $\sigma(\tilde{X}) = \sigma(\bar{X})$ can be our choice according to Lemma 5.1 and Lemma 5.2. This is because any $\tilde{X}$ that satisfies the above conditions gives $\mathbb{E}(Y|\tilde{X}) = \mathbb{E}(Y|\bar{X})$. For both theoretical and computational convenience, we fix our choice to be $\bar{X}$ from now on.

Remark 5.1 (Application of the optimal fair representation characterization to algorithm design) *In theory, we should always take $\bar{X}$ because we prove that $\bar{X}$ generates the finest sigma-algebra among all the admissible $\tilde{X}$ that is independent of Z. Especially when working with data sets with clear high-dimensional structure such as image data, one should apply more complicated models to estimate the optimal transport map instead of using affine maps. But when working with data with less high-dimensional structure such as tabular data, we hope to take advantage of the simplicity, robustness, and interpretability of linear maps in practice and hence restrict the admissible transport maps to be affine, as mentioned in Remark 3.1. Therefore, we showed that the pseudo-barycenter $X^{\dagger}$, which is equal to $\bar{X}$ in the Gaussian case and solves a relaxed version of the barycenter problem in the general distribution case, can be achieved using optimal affine maps. As a result, we apply $X^{\dagger}$ in the algorithm design and experiments. Still, if there is no concern about over-fitting or computational cost, it is recommended for strict statistical parity guarantee purposes to compute $\bar{X}$ to improve the result.*

Now, it remains to find $\bar{Y}$ to obtain the optimal fair data representation characterized by Lemma 5.3. In general, it is difficult to find $\overline{\mathbb{E}(Y|\bar{X}, Z)}$, not to mention find a $\tilde{Y}$ satisfying $\mathbb{E}(\tilde{Y}|\bar{X}) = \overline{\mathbb{E}(Y|\bar{X}, Z)}$. The key observation here is that if the Brenier's maps

$\{T_{y|\bar{X}}(\cdot, z)\}_z$ that push $\{\mathbb{E}(Y_z|\bar{X})\}_z$ forward to $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ are affine, then a straight-forward choice in $\bar{Y}$ is $\{T_{y|\bar{X}}(Y_z, z)\}_{z \in \mathcal{Z}} = T_{y|\bar{X}}(Y, Z)$. This step is the key to circumvent the post-processing nature. Therefore, following the same derivation of (41) from (1) in Section 3 to guarantee feasibility of affine maps, we relax the fairness constraint to the first two moments in Problem 3, and show a pseudo-barycenter pair provides us an exact solution to Problem 3 in the Gaussian marginal case and the optimal affine estimation in the general marginal case.

5.2 Fairness with Gaussian Marginals

Assume $\{(X_z, Y_z)\}_z$ to be non-degenerate Gaussian vectors λ-a.e. and define the following:

Definition 5.1 (Independent pseudo-barycenter: $X^\dagger$)

$$X^\dagger := T_x(X, Z), \tag{62}$$

where

$$T_x(\cdot, z) := \Sigma_{X_z}^{-\frac{1}{2}}(\Sigma_{X_z}^{\frac{1}{2}}\Sigma\Sigma_{X_z}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{X_z}^{-\frac{1}{2}} \tag{63}$$

and Σ is the unique solution to

$$\int_{\mathcal{Z}} (\Sigma^{\frac{1}{2}}\Sigma_{X_z}\Sigma^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) = \Sigma. \tag{64}$$

Definition 5.2 (Dependent pseudo-barycenter: $Y^\dagger$)

$$Y^\dagger := T_{y|X^\dagger}(Y, Z) \tag{65}$$

where

$$T_{y|X^\dagger}(\cdot, z) := \Sigma_{Y_z|X^\dagger}^{-\frac{1}{2}}(\Sigma_{Y_z|X^\dagger}^{\frac{1}{2}}\Sigma\Sigma_{Y_z|X^\dagger}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{Y_z|X^\dagger}^{-\frac{1}{2}} \tag{66}$$

with $\Sigma_{Y_z|X^\dagger} := \Sigma_{Y_zX^\dagger}\Sigma_{X^\dagger}^{-1}\Sigma_{Y_zX^\dagger}^T$, and Σ is the unique solution to

$$\int_{\mathcal{Z}} (\Sigma^{\frac{1}{2}}\Sigma_{Y_z|X^\dagger}\Sigma^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) = \Sigma \tag{67}$$

Here, to obtain (an estimation of) the solution to equations (67) and (64), we apply the iterative method (33) in Remark 2.6 when designing our algorithm in Section 6.

Since it is a direct result of Lemma 2.3 that $X^\dagger = \bar{X}$, the goal is now to show that

$$\mathbb{E}(Y^\dagger|\bar{X}) = \overline{\mathbb{E}(Y|\bar{X}, Z)}, \tag{68}$$

and therefore by Lemma 5.3 to conclude $\mathbb{E}(Y^\dagger|X^\dagger) = \mathbb{E}(Y^\dagger|\bar{X})$ indeed minimizes the estimation error while staying independent of Z.

To prove the above equation and justify the definition of the pseudo-barycenter, we need the following results: (1) existence and uniqueness of both $\bar{X}$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$; (2) affinity of the corresponding Brenier's maps $T_x(\cdot, z)$ and $T_{y|X^\dagger}(\cdot, z)$. By assumption, we have $\{\mathcal{L}(X_z)\}_z \subset \mathcal{P}_{2,ac}(\mathcal{X})$, and $\{\mathcal{L}(\mathbb{E}(Y_z|\bar{X}))\}_z \subset \mathcal{P}_{2,ac}(\mathcal{Y})$. The existence and uniqueness then follow directly from Lemma 2.1. It remains to show that the corresponding Brenier's maps are affine. But by Lemma 2.3, if $\{X_z\}_z$ and $\{\mathbb{E}(Y_z|\bar{X})\}_z$ both are from some location-scale family, then the barycenters are also from the corresponding location-scale family and the Brenier's maps are affine.

The following result shows that if $\{Y_z\}_z$ come from the same location-scale family, then $\{\mathbb{E}(Y_z|\bar{X})\}_z$ also belongs to the same location-scale family.

Lemma 5.4 (Conditional expectation preserves location-scale family) *Assume that $\{Y_z\}_z \subset \mathcal{F}(P_0)$ for some P_0, then $\{\mathbb{E}(Y_z|\bar{X})\}_z \subset \mathcal{F}(\mathcal{L}(\mathbb{E}(Y_z|\bar{X})))$ for any z.*

Proof This follows immediately from the existence of positive definite affine transformations among $\{Y_z\}_z$, Lemma 2.2, and the linearity of conditional expectation. ■

Therefore, given $\{(X_z, Y_z)\}_z$ being Gaussian vectors, we have $\{(\bar{X}, Y_z)\}$ being Gaussian vectors, which further implies that $\{\mathbb{E}(Y_z|\bar{X})\}_z$ are Gaussian vectors by Lemma 5.4. (We note that it is not necessary to apply Lemma 5.4 to show $\{\mathbb{E}(Y_z|\bar{X})\}_z$ are Gaussian because it is a well-known result in probability theory, but the lemma becomes necessary later in the case of general marginal distributions.)

Lemma 5.5 (Solution to the optimal fair data representation in the Gaussian case) *Let $\{(X_z, Y_z)\}_z$ be Gaussian vectors satisfying $\Sigma_z \succ 0$ λ-a.e., then there exists a unique barycenter pair $(\bar{X}, \overline{\mathbb{E}(Y|\bar{X}, Z)})$ which are Gaussian vectors characterized by the covariance matrix being the unique solution to*

$$\int_{\mathcal{Z}} (\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) = \Sigma \tag{69}$$

for $S \in \{\Sigma_{X_z}, \Sigma_{Y_z|X^\dagger}\}$ respectively, where $\Sigma_{Y_z|X^\dagger} = \Sigma_{Y_z X^\dagger} \Sigma_{X^\dagger}^{-1} \Sigma_{Y_z X^\dagger}^T$. Moreover, $\{T_x(\cdot, z)\}_z$ and $\{T_{y|X^\dagger}(\cdot, z)\}_z$ which push X_z and $\mathbb{E}(Y_z|\bar{X})$ respectively to $\bar{X}$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ are affine with closed-form (63) and (66). As a result, for $\lambda -$ a.e. $z \in \mathcal{Z}$, we have

$$\overline{\mathbb{E}(Y|\bar{X}, Z)}_z = T_{y|X^\dagger}(\mathbb{E}(Y_z|T_x(X_z, z)), z) = \mathbb{E}(T_{y|X^\dagger}(Y_z, z)|T_x(X_z, z)) \tag{70}$$

Proof The existence, uniqueness, and Gaussianity of the barycenter follow from Lemma 2.3, whereas the affinity of corresponding Brenier's maps results from Lemmas 5.4 and 2.2. ■

The above result provides us a theoretical foundation to apply the affine maps $\{T_x(\cdot, z)\}_z$ and $\{T_{y|X^\dagger}(\cdot, z)\}_z$ to $\{X_z\}_z$ and $\{Y_z\}_z$ respectively as a pre-processing step before the training step.

Furthermore, notice that although $T_{y|X^\dagger}(\mathbb{E}(Y_z|\bar{X}), z) = \overline{\mathbb{E}(Y_z|\bar{X}, Z)}_z$ λ-a.e. by construction, $\{T_{y|X^\dagger}(Y_z, z)\}_z$ does not agree in general: for $z_1 \neq z_2$,

$$T_{y|X^\dagger}(Y_{z_1}, z_1) \neq T_{y|X^\dagger}(Y_{z_2}, z_2). \tag{71}$$

The pseudo-barycenter solves the disagreement by merging them directly. Despite of the differences among $\{T_{y|X^\dagger}(Y_z, z)\}_z$, the L^2 projections of them on $\sigma(\bar{X})$ agree. Therefore, a direct merging of $\{T_{y|X^\dagger}(Y_z, z)\}_z$ is simply: $T_{y|X^\dagger}(Y, Z) = Y^\dagger$. It follows:

$$
\begin{aligned}
\mathbb{E}(Y^\dagger|X^\dagger) &= \mathbb{E}(Y^\dagger|\bar{X}) = \mathbb{E}(T_{y|X^\dagger}(Y,Z)|\bar{X}) \\
&= \int_{\mathcal{Z}} \mathbb{E}(T_{y|X^\dagger}(Y_z,z)|\bar{X})d\lambda(z) \\
&= \int_{\mathcal{Z}} T_{y|X^\dagger}(\mathbb{E}(Y_z|\bar{X}),z)d\lambda(z) \\
&= \int_{\mathcal{Z}} T_{y|X^\dagger}(\mathbb{E}(Y|\bar{X},Z)_z,z)d\lambda(z) \\
&= \int_{\mathcal{Z}} \overline{\mathbb{E}(Y|\bar{X},Z)_z}d\lambda(z) = \overline{\mathbb{E}(Y|\bar{X},Z)},
\end{aligned}
$$

where the second equality follows from disintegration, the third from linearity of $T_{y|\bar{X}}$, and the forth from $\mathbb{E}(Y_z|\bar{X}) = \mathbb{E}(Y|\bar{X},Z)_z$. Therefore, we have proved a result that justifies the definition of the pseudo-barycenter:

Theorem 5.1 (Justification of $Y^\dagger$ in Gaussian case) *$(X^\dagger, Y^\dagger)$ is a solution to Problem 3*

$$\inf_{(\tilde{X},\tilde{Y})\in\mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : \tilde{X}, \mathbb{E}(\tilde{Y}|\tilde{X},Z) \perp Z\}, \tag{72}$$

if $\{(X_z, Y_z)\}_z$ are non-degenerate Gaussian vectors.

5.3 The Case of General Distribution

In practice, one should not always expect the sensitive marginal data distributions to be Gaussian, and the results we derived under the assumption of Gaussianity may not apply to the general marginal distribution case. Instead, we solve the following relaxed optimal fair data representation problem:

$$\inf_{(\tilde{X},\tilde{Y})\in\mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : m_{\tilde{X}}, m_{\tilde{Y}|\tilde{X}}, \Sigma_{\tilde{X}}, \Sigma_{\tilde{Y}|\tilde{X}} \perp Z\}, \tag{73}$$

where $m_{\tilde{Y}|\tilde{X}} := \mathbb{E}(\mathbb{E}(\tilde{Y}|\tilde{X},Z))$ and similarly for $\Sigma_{\tilde{Y}|\tilde{X}}$, to find the optimal affine estimation of the true solution to the original Problem 3. The fairness guarantee of the affine estimation is the same as mentioned in Remark 3.2.

Now, we justify the pseudo-barycenter pair $(X^\dagger, Y^\dagger)$ in the case of general distributions by proving it is a solution to the relaxed optimal fair L^2-objective supervised learning problem (73). To start, notice that $(X^\dagger, Y^\dagger) \in \mathcal{D}$ and satisfies $m_{X^\dagger}, m_{Y^\dagger|X^\dagger}, \Sigma_{X^\dagger}, \Sigma_{Y^\dagger|X^\dagger} \perp Z$ by construction and therefore is admissible.

Remark 5.2 (Finest sigma-algebra vs. most variance) *Due to the relaxation, the admissible $\tilde{X} \in \mathcal{D}|_{\mathcal{X}}$ are no longer required to be independent of Z. Furthermore, without the assumption of Gaussianity, $X^\dagger$ is no longer equal to $\bar{X}$. As a result, although one can still prove $\sigma((X,Z)) = \sigma((X^\dagger, Z))$ by following the same argument in the proof of Lemma 5.2 as in the Gaussian case, but this fact now cannot imply $\sigma(\tilde{X}) \subset \sigma(X^\dagger)$ due to the lack of independence condition. Instead, the present work shows that* $\mathrm{Var}(\tilde{X}) \leq \mathrm{Var}(X^\dagger)$ *for all admissible $\tilde{X} \in \mathcal{D}|_{\mathcal{X}}$, which in general implies $\sigma(\tilde{X}) \subset \sigma(X^\dagger)$. For example, whenever set* *inclusion forms an order between $\sigma(\tilde{X})$ and $\sigma(X^\dagger)$, then it is true that* $\mathrm{Var}(\tilde{X}) \leq \mathrm{Var}(X^\dagger)$ *implies $\sigma(\tilde{X}) \subset \sigma(X^\dagger)$. As a result, we still fix $X^\dagger$ as our optimal choice among all the admissible $\tilde{X} \in \mathcal{D}|_{\mathcal{X}}$.*

In addition, for any $\Sigma \succ 0$, define

$$T_{\Sigma,x} := \Sigma_{X_z}^{-\frac{1}{2}}(\Sigma_{X_z}^{\frac{1}{2}}\Sigma\Sigma_{X_z}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{X_z}^{-\frac{1}{2}} \tag{74}$$

$$T_{\Sigma} := \Sigma_{Y_z|X_z^\dagger}^{-\frac{1}{2}}(\Sigma_{Y_z|X_z^\dagger}^{\frac{1}{2}}\Sigma\Sigma_{Y_z|X_z^\dagger}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{Y_z|X_z^\dagger}^{-\frac{1}{2}} \tag{75}$$

where $\Sigma_{Y_z|X_z^\dagger} := \mathbb{E}((\mathbb{E}(Y_z|X_z^\dagger) - m_{Y_z})(\mathbb{E}(Y_z|X_z^\dagger) - m_{Y_z})^T)$ and $\mathbb{E}(Y_z|X_z^\dagger) := \mathbb{E}(Y|X^\dagger, Z)_z$. Now, the goal is to show $(X^\dagger, Y^\dagger)$ is indeed a solution to the relaxed problem (73), under the following two assumptions:

1 Set inclusion forms an order between $X^\dagger$ and all $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : m_{\tilde{X}}, \Sigma_{\tilde{X}} \perp Z\}$.

2 $\Sigma_{Y_z|X_z^\dagger} = \Sigma_{Y_zX_z^\dagger}\Sigma_{X_z^\dagger}^{-1}\Sigma_{Y_zX_z^\dagger}^T$.

Remark 5.3 (Applicability of the assumptions) *For the first assumption, Lemma 5.6 below guarantees that $X^\dagger$ generates the finest sigma-algebra among all the admissible sigma-algebras. In other words, for any admissible $\tilde{X}$, either it generates a coarser sigma-algebra than $\sigma(X^\dagger)$ or the two sigma-algebras do not contain each other. In other words, there is no admissible $\tilde{X}$ such that $\sigma(X^\dagger) \subset \sigma(\tilde{X})$.*

The second assumption allows us to directly compute the covariance matrix of $\mathbb{E}(Y_z|X_z^\dagger)$ *from $\Sigma_{Y_zX_z^\dagger}$ and $\Sigma_{X_z^\dagger}$. The second assumption is necessary to keep our pre-processing approach. In general,* $\mathbb{E}(Y_z|X_z^\dagger)$ *is not a linear function of $X_z^\dagger$ as in the Gaussian case. When the second assumption is not true, our pre-processing approach uses $\Sigma_{Y_zX_z^\dagger}\Sigma_{X_z^\dagger}^{-1}\Sigma_{Y_zX_z^\dagger}^T$ as our best affine estimate of $\Sigma_{Y_z|X_z^\dagger}$.*

To that end, we need the following result on the relationship among the variance of the original distribution, the variance of the barycenter, and the Wasserstein distance.

Lemma 5.6 (Variance reduction of Wasserstein barycenter [39]) *Given X satisfies $\{\mathcal{L}(X_z)\}_z \subset \mathcal{P}_{2,ac}(\mathcal{X})$ and $\bar{X}$ satisfies $\mathcal{L}(\bar{X})$ being the Wasserstein barycenter of $\{\mathcal{L}(X_z)\}$, it follows that*

$$||X - \mathbb{E}(X)||_2^2 - ||\bar{X} - \mathbb{E}(\bar{X})||_2^2 = \int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(X_z), \mathcal{L}(\bar{X}))d\lambda(z) \tag{76}$$

As a result, we obtain the following:

Lemma 5.7 ($X^\dagger$ Contains the largest variance among admissible) *$X^\dagger$ is the unique solution to*

$$\sup_{\tilde{X} \in \mathcal{D}|_{\mathcal{X}}} \{\mathrm{Var}(\tilde{X}) : m_{\tilde{X}}, \Sigma_{\tilde{X}} \perp Z\}. \tag{77}$$

Proof To simplify notation, by the invariance of variance under translation and Lemma 2.1, we can assume without loss of generality that $m_{X_z} = 0$ $\lambda - a.e.$ in the rest of the proof, which only deal with variance and Wasserstein distance. Now, for $\lambda - a.e.$ $z \in \mathcal{Z}$, we have

$$\begin{aligned}||X_z - T_{\Sigma,x}(X_z, z)||_2^2 =&||X_z||_2^2 + ||T_{\Sigma,x}(X_z, z)||_2^2 - 2\langle X_z, T_{\Sigma,x}(X_z, z)\rangle_2\\ =&\,\text{Trace}(\Sigma_{X_z}) + \text{Trace}(\Sigma) - 2\mathbb{E}(X_z^T T_{\Sigma,x}(X_z, z))\\ =&\,\text{Trace}(\Sigma_{X_z}) + \text{Trace}(\Sigma) - 2\langle T_{\Sigma,x}, \Sigma_{X_z}\rangle_F\\ =&\,\text{Trace}(\Sigma_{X_z'}) + \text{Trace}(\Sigma) - 2\langle T_{\Sigma,x}, \Sigma_{X_z'}\rangle_F\\ =&||X_z' - T_{\Sigma,x}(X_z', z)||_2^2\\ =&\mathcal{W}_2^2(\mathcal{L}(X_z'), \mathcal{L}(T_{\Sigma,x}(X_z')))\end{aligned}$$

where $X' \sim \mathcal{N}(m_X, \Sigma_X)$ is the Gaussian analog of X and $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product.

Similarly, by the disintegration theorem, we also have for $S \in \{X, X^\dagger\}$

$$\text{Var}(S) = ||S||_2^2 = \int_{\mathcal{Z}} ||S_z||_2^2 d\lambda = \int_{\mathcal{Z}} \text{Trace}(\Sigma_{S_z}) d\lambda. \tag{78}$$

Therefore, it follows from Lemma 5.6 that

$$\begin{aligned}\text{Var}(X) - \text{Var}(X^\dagger) =&\,\text{Var}(X') - \text{Var}((X')^\dagger)\\ =&\,\text{Var}(X') - \text{Var}(\bar{X}')\\ =&\int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(X_z'), \mathcal{L}(\bar{X}'))d\lambda(z).\end{aligned}$$

Finally, assume there exists a $\tilde{X} \in \mathcal{D}|_{\mathcal{X}}$ such that $\text{Var}(X^\dagger) \leq \text{Var}(\tilde{X})$. It follows $\text{Var}(X') - \text{Var}(\tilde{X}') \leq \text{Var}(X') - \text{Var}((X')^\dagger) = \text{Var}(X') - \text{Var}(\bar{X}')$. But since $m_{\tilde{X}'}, \Sigma_{\tilde{X}'} \perp Z$, we have $\tilde{X}' \perp Z$ as $\tilde{X}'$ is Gaussian by construction. In other words, there exists a $\tilde{X}' \perp Z$ such that

$$\int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(X_z'), \mathcal{L}(\tilde{X}'))d\lambda(z) \leq \int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(X_z'), \mathcal{L}(\bar{X}'))d\lambda(z) \tag{79}$$

which contradicts the uniqueness of $\bar{X}'$. ■

The above lemma shows that $\text{Var}(\tilde{X}) \leq \text{Var}(X^\dagger)$ for all admissible $\tilde{X} \in \mathcal{D}|_{\mathcal{X}}$ satisfies $m_{\tilde{X}}, \Sigma_{\tilde{X}} \perp Z$, which together with the first assumption imply $\sigma(\tilde{X}) \subset \sigma(\bar{X})$ in practice. Therefore, from now on, we fix the choice of $\tilde{X}$ to be $X^\dagger$ and prove the general characterization result based on the two assumptions listed above.

It remains to justify the choice of $Y^\dagger$. To do so, we need the following lemma, which provides a multi-marginal characterization of the optimal affine map.

Lemma 5.8 (Projection Lemma for conditional expectations) *Given* $m_{Y_z|X_z^\dagger} = 0$ *and* $\Sigma_{Y_z|X_z^\dagger} \succ 0$ *λ-a.e., for any* $\Sigma \succ 0$,

$$\inf_{\mathbb{E}(\tilde{Y}|X^\dagger):\Sigma_{\tilde{Y}_z|X_z^\dagger}=\Sigma} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(\mathbb{E}(Y_z|X_z^\dagger)), \mathcal{L}(\mathbb{E}(\tilde{Y}_z|X_z^\dagger)))d\lambda(z) \tag{80}$$

admits a unique solution, denoted by $Y_\Sigma^\dagger$, *that has the form*

$$Y_\Sigma^\dagger := T_\Sigma(Y, Z) \tag{81}$$

where $T_\Sigma(\cdot, z) := \Sigma_{\tilde{Y}_z|X_z^\dagger}^{-\frac{1}{2}}(\Sigma_{\tilde{Y}_z|X_z^\dagger}^{\frac{1}{2}}\Sigma\Sigma_{\tilde{Y}_z|X_z^\dagger}^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_{\tilde{Y}_z|X_z^\dagger}^{-\frac{1}{2}}$

Proof This is a direct corollary from Lemma 3.3. ■

Finally, we are ready to prove the justification of the pseudo-barycenter in the case of general distributions.

Theorem 5.2 (Justification of $(X^\dagger, Y^\dagger)$ in general distribution case) $\mathbb{E}(Y^\dagger|X^\dagger)$ *is a solution to*

$$\inf_{(\tilde{X},\tilde{Y})\in\mathcal{D}} \{||Y - \mathbb{E}(\tilde{Y}|\tilde{X})||_2^2 : m_{\tilde{X}}, m_{\tilde{Y}|\tilde{X}}, \Sigma_{\tilde{X}}, \Sigma_{\tilde{Y}|\tilde{X}} \perp Z\} \tag{82}$$

under the assumptions: (1) set inclusion forms an order between $X^\dagger$ *and all* $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : m_{\tilde{X}}, \Sigma_{\tilde{X}} \perp Z\}$; *and (2)* $\Sigma_{Y_z|X_z^\dagger} = \Sigma_{Y_zX_z^\dagger}\Sigma_{X_z^\dagger}^{-1}\Sigma_{Y_zX_z^\dagger}^T$.

Proof The choice of $X^\dagger$ follows from the first assumption and Lemma 5.7. It remains to show that $Y^\dagger$ is a solution to

$$\inf_{\tilde{Y}\in\mathcal{D}|_{\mathcal{Y}}} \{||Y - \mathbb{E}(\tilde{Y}|X^\dagger)||_2^2 : m_{\tilde{Y}|X^\dagger}, \Sigma_{\tilde{Y}|X^\dagger} \perp Z\} \tag{83}$$

Fix $\Sigma \succ 0$ arbitrary, we have

$$||Y - \mathbb{E}(Y_\Sigma^\dagger|X^\dagger)||_2^2 - ||Y - \mathbb{E}(Y|X^\dagger)||_2^2 = \int_{\mathcal{Z}} ||\mathbb{E}(Y_z - Y_{\Sigma,z}^\dagger|X_z^\dagger)||_2^2 d\lambda(z) \tag{84}$$

and it follows from Lemma 5.8 that

$$\begin{aligned}\int_{\mathcal{Z}} ||\mathbb{E}(Y_z - Y_{\Sigma,z}^\dagger|X_z^\dagger)||_2^2 d\lambda(z) &= \int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(\mathbb{E}(Y_z|X_z^\dagger)), \mathcal{L}(T_\Sigma(\mathbb{E}(Y_z|X_z^\dagger), z))d\lambda(z) \\ &= \min_{\nu:\Sigma_{\nu_z}=\Sigma} \int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(\mathbb{E}(Y_z|X_z^\dagger)), \nu_z)d\lambda(z)\end{aligned}$$

Therefore, (73) boils down to the following:

$$\inf_{\Sigma\succ 0}\{\int_{\mathcal{Z}} ||\mathbb{E}(Y_z - Y_{\Sigma,z}^\dagger|X_z^\dagger)||_2^2 d\lambda(z)\}. \tag{85}$$

Finally, notice that

$$\begin{aligned}
&\int_{\mathcal{Z}} ||\mathbb{E}(Y_z - Y^{\dagger}_{\Sigma,z}|X^{\dagger}_z)||_2^2 d\lambda(z) \\
=& \int_{\mathcal{Z}} ||\mathbb{E}(Y_z|X^{\dagger}_z) - T_{\Sigma}(\mathbb{E}(Y_z|X^{\dagger}_z), z)||_2^2 d\lambda(z) \\
=& \int_{\mathcal{Z}} ||\mathbb{E}(Y_z|X^{\dagger}_z)||_2^2 + ||T_{\Sigma}(\mathbb{E}(Y_z|X^{\dagger}_z), z)||_2^2 - 2\langle \mathbb{E}(Y_z|X^{\dagger}_z), T_{\Sigma}(\mathbb{E}(Y_z|X^{\dagger}_z), z)\rangle_2 d\lambda(z) \\
=& \int_{\mathcal{Z}} \text{Trace}(\Sigma_{Y_z|X^{\dagger}_z}) + \text{Trace}(\Sigma) - 2\mathbb{E}(\mathbb{E}(Y_z|X^{\dagger}_z)^T T_{\Sigma}(\mathbb{E}(Y_z|X^{\dagger}_z), z)) d\lambda(z) \\
=& \int_{\mathcal{Z}} \text{Trace}(\Sigma_{Y_z|X^{\dagger}_z}) + \text{Trace}(\Sigma) - 2\langle T_{\Sigma}, \Sigma_{Y_z|X^{\dagger}_z}\rangle_F d\lambda(z) \\
=& \int_{\mathcal{Z}} ||\mathbb{E}(Y_z|X^{\dagger}_z)' - T_{\Sigma}(\mathbb{E}(Y_z|X^{\dagger}_z)', z)||_2^2 d\lambda(z)
\end{aligned}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product and $X' \sim \mathcal{N}(m_X, \Sigma_X)$ denotes the Gaussian analog of X. It follows from the definition of $Y^{\dagger}$ and Lemma 2.3 that $\int_{\mathcal{Z}} ||\mathbb{E}(Y_z - Y^{\dagger}_z|X^{\dagger})||_2^2 d\lambda(z)$ is the lower bound of (85). The proof is complete. ■

To conclude, given an arbitrary L^2-objective supervised learning model that aims to estimate conditional expectation, the training via $(X^{\dagger}, Y^{\dagger})$ results in an estimate of $\overline{\mathbb{E}(Y|\bar{X}, Z)}$. In other words, any supervised learning model trained via $(X^{\dagger}, Y^{\dagger})$ is guaranteed to be independent of Z in the location-scale family marginal case (or, to have first two moments independent of Z in the general marginal case), while resulting in the minimum prediction error among all the admissible functions of some specific model due to the training step. Here, the assumption is that the test sample distribution is the same as the training sample distribution, which is a ubiquitous assumption for machine learning.

5.4 Optimal Fair Data Representation at the Pareto Frontier

Finally, we extend the pseudo-barycenter pair, which is the solution to the optimal fair data representation, to the fair data representation at the Pareto frontier using McCann interpolation via a similar approach as we derived the post-processing Pareto frontier in Section 4. But notice a direct application of Theorem 4.1 does not work here because there is no direct interpolation between $E(Y|X,Z)$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ due to the change of the underlying sigma-algebra. Therefore, we apply a diagonal argument, Remark 5.4, to estimate the interpolation between $E(Y|X,Z)$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ and thus the fair data representation at the Pareto frontier.

To start, we derive the following post-processing optimal trade-off result directly from Theorem 4.1 for a fixed choice of $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : \tilde{X} \perp Z\}$. For any $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, define $L_{y|\tilde{X}}$, $D_{y|\tilde{X}}$, and $V_{y|\tilde{X}}$ as follows:

$$L_{y|\tilde{X}}(f(\tilde{X}, Z)) := (\int_{\mathcal{Z}} ||\mathbb{E}(Y_z|\tilde{X}) - f(\tilde{X}, Z)_z||_2^2 d\lambda(z))^{\frac{1}{2}} \tag{86}$$

$$D_{y|\tilde{X}}(f(\tilde{X}, Z)) := (\int_{\mathcal{Z}^2} \mathcal{W}_2^2(f(\tilde{X}, Z)_{z_1}, f(\tilde{X}, Z)_{z_2}) d\lambda(z_1) d\lambda(z_2))^{\frac{1}{2}}. \tag{87}$$

To simplify notation, for any $T' : \mathcal{Y} \times \mathcal{Z} \to \mathcal{Y}$, we also define the following:

$$L_{y|\tilde{X}}(T') := (\int_{\mathcal{Z}} ||\mathbb{E}(Y_z|\tilde{X}) - T'_z(\mathbb{E}(Y_z|\tilde{X}))||_2^2 d\lambda(z))^{\frac{1}{2}} \tag{88}$$

$$D_{y|\tilde{X}}(T') := (\int_{\mathcal{Z}^2} \mathcal{W}_2^2((T'_{z_1})_\sharp \mathcal{L}(\mathbb{E}(Y_{z_1}|\tilde{X})), (T'_{z_2})_\sharp \mathcal{L}(\mathbb{E}(Y_{z_2}|\tilde{X}) d\lambda(z_1) d\lambda(z_2))^{\frac{1}{2}}. \tag{89}$$

Also, let T denote the optimal transport map from $\{\mathbb{E}(Y_z|\tilde{X})\}_z$ to the barycenter $\overline{\mathbb{E}(Y|\tilde{X}, Z)}$, let $T(t), t \in [0, 1]$ be the McCann interpolation, and define

$$\begin{aligned} V_{y|\tilde{X}} := L_{y|\tilde{X}}(T) &= (\int_{\mathcal{Z}} ||\mathbb{E}(Y_z|\tilde{X}) - T_z(\mathbb{E}(Y_z|\tilde{X}))||_2^2 d\lambda(z))^{\frac{1}{2}} && (90) \\ &= (\int_{\mathcal{Z}} ||\mathbb{E}(Y_z|\tilde{X}) - \overline{\mathbb{E}(Y|\tilde{X}, Z)}||_2^2 d\lambda(z))^{\frac{1}{2}}. && (91) \end{aligned}$$

Then the result below follows directly similar to the proof of Theorem 4.1.

Corollary 5.1 (Pareto frontier for conditional expectation on fixed sigma-algebra) *Given $L_{y|\tilde{X}}$, $D_{y|\tilde{X}}$, and $V_{y|\tilde{X}}$ defined above, we have*

$$V_{y|\tilde{X}} \leq L_{y|\tilde{X}}(f(\tilde{X}, Z)) + \frac{1}{\sqrt{2}} D_{y|\tilde{X}}(f(\tilde{X}, Z)) \tag{92}$$

where equality holds if and only if $f(\tilde{X}, z) = T(t)(\mathbb{E}(Y_z|\tilde{X}), z)$ λ-a.e. for $t \in [0, 1]$ as

$$L_{y|\tilde{X}}(T(t)) = t L_{y|\tilde{X}}(T(0)) = t V_{y|\tilde{X}}, \tag{93}$$

$$\frac{1}{\sqrt{2}} D_{y|\tilde{X}}(T(t)) = \frac{1}{\sqrt{2}}(1 - t) D_{y|\tilde{X}}(T(0)) = (1 - t) V_{y|\tilde{X}}. \tag{94}$$

The above result shows that by fixing $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : \tilde{X} \perp Z\}$, the McCann interpolation between Id and $T_{y|\tilde{X}}$ yields the Pareto frontier from $\mathbb{E}(Y|\tilde{X}, Z)$ to $\overline{\mathbb{E}(Y|\tilde{X}, Z)}$, which is a weak version of the true frontier from $\mathbb{E}(Y|X, Z)$ to $\overline{\mathbb{E}(Y|\tilde{X}, Z)}$. The only difficulty remaining is to coarsen the underlying sigma-algebra from $\sigma(X, Z)$ to $\sigma(\bar{X})$. But by Remark 5.2, we know that one can coarsen the sigma-algebra by reducing the variance. Therefore, we apply a diagonal argument to estimate the McCann interpolation between (X, Y) and $(\bar{X}, \bar{Y})$.

Remark 5.4 (Diagonal estimate of the post-processing Pareto frontier) *The key observation is that the optimal affine transport map that pushes (X, Y) forward to $(X^\dagger, Y^\dagger)$ is the pair $(T_x, T_{y|\bar{X}})$. Therefore, McCann interpolation between Id and T_x can optimally reduce variance and thereby coarsen $\sigma((X, Z))$ to $\sigma(X^\dagger)$, whereas the interpolation betwen Id and $T_{y|\bar{X}}$ forms an estimation of the geodesic path between Y and $Y^\dagger$. Therefore, the present work matches the two interpolations diagonally*

$$(T_x(t), T_{y|\bar{X}}(t)) := ((1 - t) Id_x + t T_x, (1 - t) Id_y + t T_{y|\bar{X}}),$$

to estimate the true optimal fair data representation at the Pareto frontier.

Finally, since $X^\dagger$ and $\mathbb{E}(Y^\dagger|X^\dagger)$ are the estimation of $\bar{X}$ and $\overline{\mathbb{E}(Y|\bar{X}, Z)}$, respectively, as shown in the last section, it follows from Corollary 5.1 and Remark 5.4 that

$$\mathbb{E}(T_{y|\bar{X}}(t)(Y, Z)|T_x(t)(X, Z)), t \in [0, 1] \tag{95}$$

provides a pre-processing estimate of the Pareto frontier from $\mathbb{E}(Y|X, Z)$ to $\overline{\mathbb{E}(Y|\bar{X}, Z)}$ that is characterized by Theorem 4.1.

6. Algorithm Design

In this section, we propose two algorithms based on the theoretical results above. Algorithm 2 is designed for the fair learning outcome in the post-processing approach and for the dependent variable in fair data representation, whereas Algorithm 1 is designed for the independent variable in fair data representation.

1. For practitioners who want to generate fair learning outcomes along the Pareto frontier, Algorithm 2 takes the learning outcomes marginals $\{f(X, Z)_z\}_z$ as input and outputs the learning outcomes at (the optimal affine estimation of) the post-processing estimation of the Pareto frontier: $\{f(X, Z)(t)\}_{t \in [0,1]}$, which is the Wasserstein geodesic paths from the original learning outcome, $f(X, Z)(0)$, to the estimate of the optimal fair learning outcome, $f(X, Z)(1)$. Here, $f(X, Z)(1)$ is the best estimate of the optima fair learning outcome based on the provided learning outcome $\{f(X, Z)_z\}_z$.
2. For practitioners who want to generate a fair data representation, Algorithm 1 and Algorithm 2 take in respectively the marginal independent and dependent data: $\{X_z\}_z$ and $\{Y_z\}_z$, then outputs respectively the independent and dependent data representations along the Wasserstein geodesics from the marginals to their pseudo-barycenter: $\{(X^\dagger(t), Y^\dagger(t))\}_{t \in [0,1]}$. So that any conditional expectation estimation supervised learning model trained via $\{(X^\dagger(t), Y^\dagger(t))\}_{t \in [0,1]}$ results in (an diagonal affine estimation of) the learning outcome at the Pareto frontier.

The choice of the Frobenius norm in Step 1 is due to computational efficiency. Any matrix norm would work.

Remark 6.1 (Solution to alternative fair data representation constraint) *In Section 1.3, the present work shows two alternative fair data representation constraints: (1) $(\tilde{X}, \tilde{Y}) \perp Z$ and (2) $\tilde{X} \perp Z$, which offer different trade-offs between fairness protection and utility. If a practitioner applies the alternative constraint, the proposed algorithms can be applied to generate (the optimal affine estimation of) corresponding fair data representation as the following:*

1 For $(\tilde{X}, \tilde{Y}) \perp Z$, one applies Algorithm 1 to both $\{(X_z, Y_z)\}_z$. This alternative is especially useful when practitioners or data publishers do not know which features would be chosen as independent or dependent.

2 For $\tilde{X} \perp Z$, one applies Algorithm 1 to $\{X_z\}_z$ and leaves $\{Y_z\}$ untouched.

Algorithm 1: Pseudo-Barycenter Geodesics for Independent Variable

```
Input: marginal data sets $\{X_z\}_z$, stop criterion $\epsilon$;
Step 1: Find the optimal barycenter covariance;
Initialization: $\delta = \infty$, $\Sigma = rand$ or $Id$
while $\delta > \epsilon$ do
    $\Sigma_{new} = \frac{1}{|X|} \sum_z |X_z| (\Sigma^{\frac{1}{2}} \Sigma_{X_z} \Sigma^{\frac{1}{2}})^{\frac{1}{2}}$;        // (33)
    $\delta = ||\Sigma - \Sigma_{new}||_F$;
    $\Sigma = \Sigma_{new}$;
end
Step 2: Find the optimal affine transport maps;
$T_z = \Sigma_{X_z}^{-\frac{1}{2}} (\Sigma_{X_z}^{\frac{1}{2}} \Sigma \Sigma_{X_z}^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_{X_z}^{-\frac{1}{2}}$;        // (63)
Step 3: Find the geodesic path to independent pseudo-barycenter;
$X_z^\dagger(t) = T_z(t)(X_z - m_{X_z}) + m_X$;        // (62)
where $T_z(t) := (1-t)Id + tT_z$, $t \in [0,1]$;        // (55)
Step 4 (optional): For binary rows $X_{i \in I}$, reshape $(X^\dagger(t))_i$ to binary by randomized rounding for all $i \in I$;
For all $X_i$ binary: $p(t) = \frac{(X_z^\dagger(t))_i}{\max((X_z^\dagger(t))_i) - \min((X_z^\dagger(t))_i)}$, $(X_z^\dagger(t))_i \sim \text{Bernoulli}(p(t))$;
Step 5 (optional): If sensitive information needs to be attached, merge the marginals back with mitigating $Z$;
$X_z^\dagger(t) = (X_z(t), z(t))$ where $z(t) = (1-t)(z - m_Z) + m_Z$, $t \in [0,1]$
Output: $\{\{X_z^\dagger(t)\}_{z \in \mathcal{Z}}\}_{t \in [0,1]}$
```

```
Algorithm 2: Dependent (or Post-processing) Pseudo-Barycenter Geodesics
  Input: marginal data sets $\{Y_z\}_z$ (post-processing: $\{f(X,Z)_z\}_z$), stop criterion $\epsilon$;
  Step 1: Find the optimal barycenter covariance;
  Initialization: $\delta = \infty$, $\Sigma = rand$ or $Id$
  while $\delta > \epsilon$ do
      $\Sigma_{new} = \frac{1}{|Y|}\sum_z |Y_z|(\Sigma^{\frac{1}{2}}\Sigma_{Y_z|X_z^\dagger}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}$    // (33)
      (post-processing: $\Sigma_{new} = \frac{1}{|Y|}\sum_z |f(X,Z)_z|(\Sigma^{\frac{1}{2}}\Sigma_{f(X,Z)_z}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}$);
      $\delta = ||\Sigma - \Sigma_{new}||_F$;
      $\Sigma = \Sigma_{new}$;
  end
  Step 2: Find the optimal affine transport maps;
 $T_z = \Sigma^{-\frac{1}{2}}_{Y_z|X_z^\dagger}(\Sigma^{\frac{1}{2}}_{Y_z|X_z^\dagger}\Sigma\Sigma^{\frac{1}{2}}_{Y_z|X_z^\dagger})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}_{Y_z|X_z^\dagger}$    // (66)
  (post-processing: $T_z = \Sigma^{-\frac{1}{2}}_{f(X,Z)_z}(\Sigma^{\frac{1}{2}}_{f(X,Z)_z}\Sigma\Sigma^{\frac{1}{2}}_{f(X,Z)_z})^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}_{f(X,Z)_z}$);    // (38)
   Step 3: Find the geodesic path to dependent pseudo-barycenter;
 $Y_z^\dagger(t) = T_z(t)(Y_z - m_{Y_z}) + m_Y$    // (65)
 where $T_z(t) := (1-t)Id + tT_z, t \in [0,1]$    // (55)
 (post-processing: $f(X,Z)_z(t) = T_z(t)(f(X,Z)_z - m_{f(X,Z)_z}) + m_{f(X,Z)}$);    // (37)
  Step 4 (optional): For binary rows $Y_{i\in I}$ (post-processing: $(f(X,Z))_{i\in I}$), reshape
    $(Y^\dagger(t))_i$ (post-processing: $(f(X,Z)(t))_{i\in I}$) to binary by randomized rounding for
    all $i \in I$;
 For all $Y_i$ binary: $p(t) = \frac{(Y_z^\dagger(t))_i}{\max((Y_z^\dagger(t))_i) - \min((Y_z^\dagger(t))_i)}$, $(Y_z^\dagger(t))_i \sim$ Bernoulli$(p(t))$;
  Output: $\{\{Y_z^\dagger(t)\}_{z\in\mathcal{Z}}\}_{t\in[0,1]}$ (post-processing: $\{\{f(X,Z)_z(t)\}_{z\in\mathcal{Z}}\}_{t\in[0,1]}$)
```

7. Empirical Study: Fair Supervised Learning

In this section, we present numerical experiments with the proposed Algorithms 1 and 2 from Section 6. The proposed fair data representation method is bench-marked against two baselines:

1. the prediction model trained via the original data (denoted by "supervised learning name" in the experiment result figure below): supervised learning models trained via data including the sensitive variable provide an estimation of statistical disparity resulting from both disparate treatment and impact.
2. the prediction model trained via data excluding the sensitive variable (denoted by "supervised learning name + Excluding Z"): supervised learning models trained via data excluding the sensitive variable provide an estimation of statistical disparity resulting from only disparate impact.

7.1 Benchmark Data and Comparison Methods

For comparison, we implement the following known methods for different types of supervised learning tests:

1. For classification test, the present work compares the current state-of-the-art pre-processing methods [13, 44] ("supervised learning name + Calmon or Zemel", the later is also known as "Learning Fair Representation") with the proposed fair data representation methods ("supervised learning name + pre-proc. Pareto frontier Est. or Pseudo-barycenter").
2. For uni-variate regression test, we compare the post-processing Wasserstein barycenter based fair regression [18] ("supervised learning name + Chzhen") with the proposed post-processing pseudo-barycenter methods ("supervised learning name + post-proc. Pareto frontier Est. or Pseudo-barycenter") and the fair data representation methods.
3. For multi-variate supervised learning test, we compare the post-processing pseudo-barycenter methods with the fair data representation methods.

The reasons for this choice are as follows: (1) the known attempts via the pre-processing approach are only available for fair classification; (2) the post-processing Wasserstein barycenter based methods on fair classification are analogous to the one on fair regression, which is shown to outperform other in-processing or post-processing methods in reducing discrimination while preserving accuracy; (3) there exists no practical attempt along the Wasserstein characterization approach to multi-dimensional supervised learning due to the computational complexity of finding the barycenter and the optimal transport maps.

We adopt the following metrics of accuracy and discrimination that are frequently used in fair machine learning experiments on various data sets: (1) For fair classification, the prediction accuracy, and statistical disparity are quantified respectively by AUC (area under the Receiver Operator Characteristic curve) and

Definition 7.1 (Classification discrimination)

$$Discrimination = \max_{z,z' \in \mathcal{Z}} \left| \frac{\mathbb{P}(\hat{Y}_z = 1)}{\mathbb{P}(\hat{Y}_{z'} = 1)} - 1 \right|$$

as defined in [13]. (2) For univariate supervised learning, the prediction error and statistical disparity are quantified respectively by MSE (mean squared error, equivalent to the squared L^2 norm on sample probability space) and KS (Kolmogorov-Smirnov) distance as in [18] for indirect comparison purpose. So that readers can compare the proposed methods indirectly with other methods that are tested in [13, 18, 44] and their references. (3) For univariate and multivariate supervised learning, the prediction error and statistical disparity are quantified respectively by L^2 and $\mathcal{W}_2$ (Wasserstein) distances, which are the quantification the current work adopts to prove the Pareto frontier in the above sections.

In addition, we perform tests on four benchmark data sets: CRIME, LSAC, Adult, COMPAS, which are also frequently used in fair learning experiments. A brief summary is given below. For all the test results, we apply 5-fold cross-validation with 50% training and 50% testing split, except for 90% training and 10% testing split in the linear regression test on LSAC due to the high computational cost of the post-processing Wasserstein barycenter method [18]. Therefore, interested readers can also compare the pseudo-barycenter test results indirectly to other methods tested in [13, 18].

Data set	Tests	Data size	dim(X)	dim(Y)
UCI Adult	logit regression, random forest	162805	16	1
COMPAS	logit regression, random forest	26390	7	1
LSAC	linear regression, ANN	20454	9	1
CRIME	linear regression, ANN	1994	97	1
CRIME	linear regression, ANN	1994	87	11

- Communities and Crime Data Set (CRIME) contains the social, economic, law executive, and judicial data of communities in the United States with 1994 examples [35]. The task of univariate learning is to predict the number of crimes per 10^5 population using the rest of the information on the data set. Here, race is the sensitive information and, for (indirect) comparison purposes, we made race a binary categorical variable of whether the percentage of the African American population (racepctblack) is greater than 30%.

  In multivariate supervised learning on CRIME, we keep the same sensitive variable. But the learning task is to predict the following vector that represents the local housing and rental market information: (low quartile occupied home value, median home value, high quartile home value, low quartile rent, median rent, high quartile rent, median gross rent, number of immigrants, median number of bedrooms, number of vacant households, number of crimes).
- LSAC National Longitudinal Bar Passage Study data set (LSAC) contains social, economic, and personal data of law school students with 20454 examples [42]. The goal of univariate models is to predict the students' GPA using other information on the data set. Here, race is the sensitive variable and, for (indirect) comparison purposes, we make it a binary variable on whether the student is non-white.
- UCI Adult Data Set (Adult) contains the 1994 Census data with 162805 examples [7]. The goal is to predict the binary categorization (whether gross annual income is

greater than 50k) using age, education years, and gender, where gender is the sensitive information.

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a benchmark set of data from Broward County, Florida for algorithmic bias studies [5]. Following [13], the goal here is to predict whether an individual would commit any violent crime while race is the sensitive binary variable (African-American and Caucasian).

7.2 Numerical Result

In this subsection, we summarize the experimental results[3].

The classification test result is summarized in Figure 3 below. Here, the vertical and horizontal axes are AUC and Discrimination defined in Definition 7.1. That is, the more upper-left, the better the result. The first row of Figure 3 shows the results of logistic regression (left) and random forest (right) on Adult whereas the second shows the corresponding results on COMPAS.

[3] The code for the results of our experiments is available online at: `github.com/xushizhou/fair_data_representation`

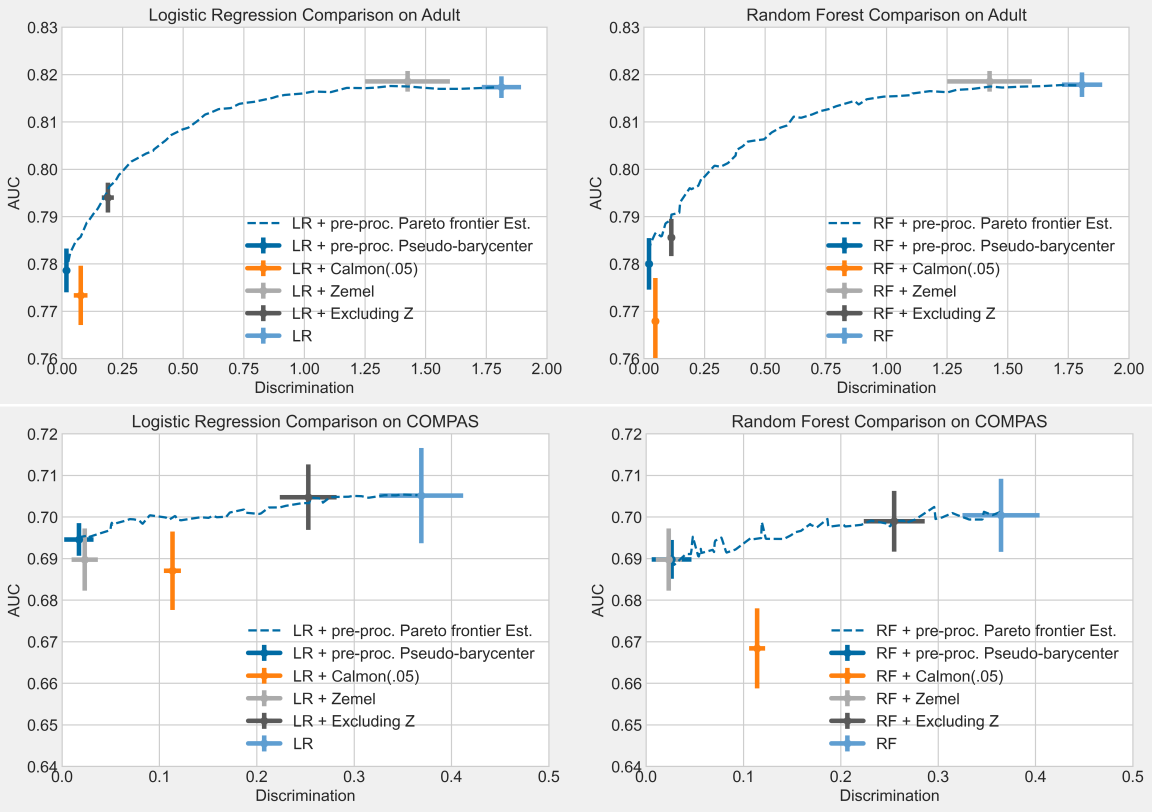


Figure 3: As shown in the classification test above, the proposed fair data representation method (+ Pre-proc. Pareto frontier Est. or Pseudo-barycenter) outperforms the other methods (+ Zemel or + Calmon) in estimating the optimal fair learning outcome. It reduces the Discrimination metric to nearly zero while keeping the relatively high level of AUC with both logistic regression (LR) and random forest (RF) on both Adult and COMPAS. Furthermore, fair data representation method offers flexibility in choosing the desired trade-off while other methods only estimate a random point near the Pareto frontier.

We note that there exists a large disparate impact in the learning outcome on COMPAS due to the relatively small difference between the “Discrimination” of learning outcome on the original data (LR and RF) and the outcome on the data excluding Z (LR and RF + Excluding Z). Therefore, a further reduction of statistical disparity is needed. In contrast, the relatively large difference in the Adult data set implies a small disparate impact. That is, a simple exclusion of the sensitive variable Z results in a significant improvement in fairness.

For further reduction of statistical disparity, it is clear from the experiment results on both COMPAS and Adult that the estimation via the Wasserstein geodesics to Pseudo-barycenter (LR and RF + Pseudo-barycenter) consistently outperforms LR and RF + Calmon by obtaining lower Discrimination with higher AUC.

In addition, although “LR and RF + Zemel” achieves a point near the Pareto frontier estimated by the proposed Pseudo-barycenter methods, the point estimation is rather random. Hence, “+ Zemel” is not consistent in estimating the optimal fair learning outcome (the end point of the Pareto curve). Practitioners cannot know which point on the Pareto frontier is estimated by “+ Zemel”. In comparison, the pseudo-barycenter methods are consistent in estimating the optimal fair learning outcome. In addition, they providef the entire Pareto frontier, and hence offer practitioners the flexibility to choose the desired trade-off. Moreover, the proposed method works for any model that aims to estimate conditional expectation, including classification and regression, while "+ Zemel" only works for classification.

The univariate regression test result on the LSAC and the one on CRIME are shown respectively in Figure 4 and 5 below. Here, the vertical and horizontal axes in the first rows are MSE and KS distance. The corresponding axes in the second row are the L^2-quantified test error and the $\mathcal{W}_2$ distance that quantifies the remaining statistical disparity among sensitive groups. Therefore, the more lower-left, the better is the result in both rows. The two supervised learning methods we use are linear regression and artificial neural networks (ANN with 4 linearly stacked layers where each of the first three layers has 32 units all with ReLu activation while the last has 1 unit with linear activation).

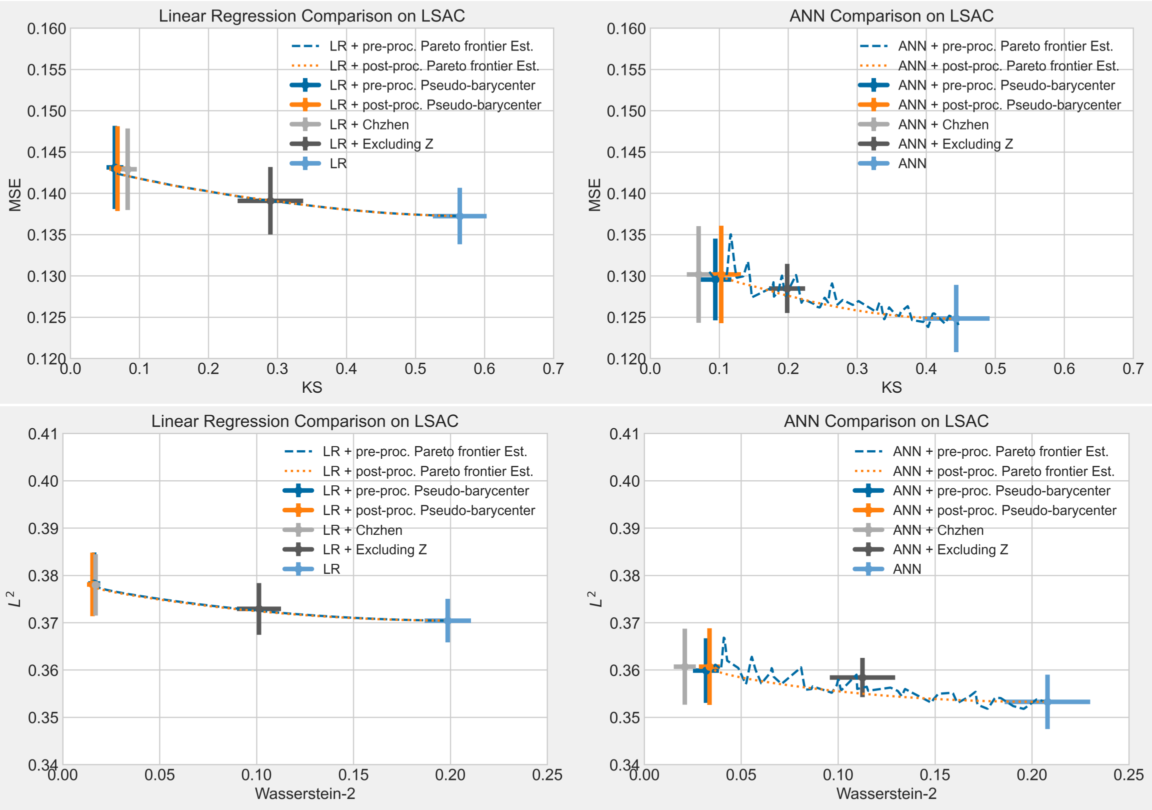


Figure 4: As shown in the univariate regression test on LSAC above, the proposed fair data representation method (+ pre-proc. Pareto frontier Est. or Pseudo-barycenter) and the post-processing pseudo-barycenter geodesics method (+ post-proc. Pareto frontier Est. or Pseudo-barycenter) achieved similar performance as the exact barycenter method (+ Chzhen). The proposed methods outperformed "+ Chzhen" with linear regression and were exceeded with the artificial neural network, both by a narrow margin. But the performance of the proposed methods is achieved at 0.0128% of the time costs "+ Chzhen" (see Figure 7 below). In addition, the proposed methods offer the flexibility of choosing the desired (optimal) trade-off between utility loss (MSE or L^2-loss) and statistical disparity (KS or $\mathcal{W}_2$ distance), whereas "+ Chzhen" only estimate the end point of the Pareto curve.

In the regression tests, post-processing Pareto frontier estimation via ANN is smooth while the pre-processing estimation is not. Here, the smoothness is due to the McCann interpolation between the identity matrix and the optimal transport map in the post-processing approach. The non-smoothness is due to the randomness in training the neural network. When testing fair data representations via ANN, one has to train the neural network for the data representation at every time $t \in [50]$. Hence, the randomness in ANN training results in the non-smoothness in the Pareto frontier estimation via fair data representations.

On the LSAC data set, the proposed methods (+ pre-proc. Pseudo-barycenter and + post-proc. Pseudo-barycenter) obtains a similar performance as the post-processing exact Wasserstein barycenter method (+ Chzhen): the proposed methods outperformed the exact method in the linear regression test and were outperformed by the exact method in the non-linear artificial neural network tests, which is consistent with our theoretical results. But the performance of the proposed methods is achieved at 0.81 seconds on average, whereas the average time cost of "+ Chzhen" is 6365.98 seconds (see Figure 7 below). In addition, we gained the flexibility in choosing the desired trade-off, computational efficiency, model selection, parameter tuning, and composition.

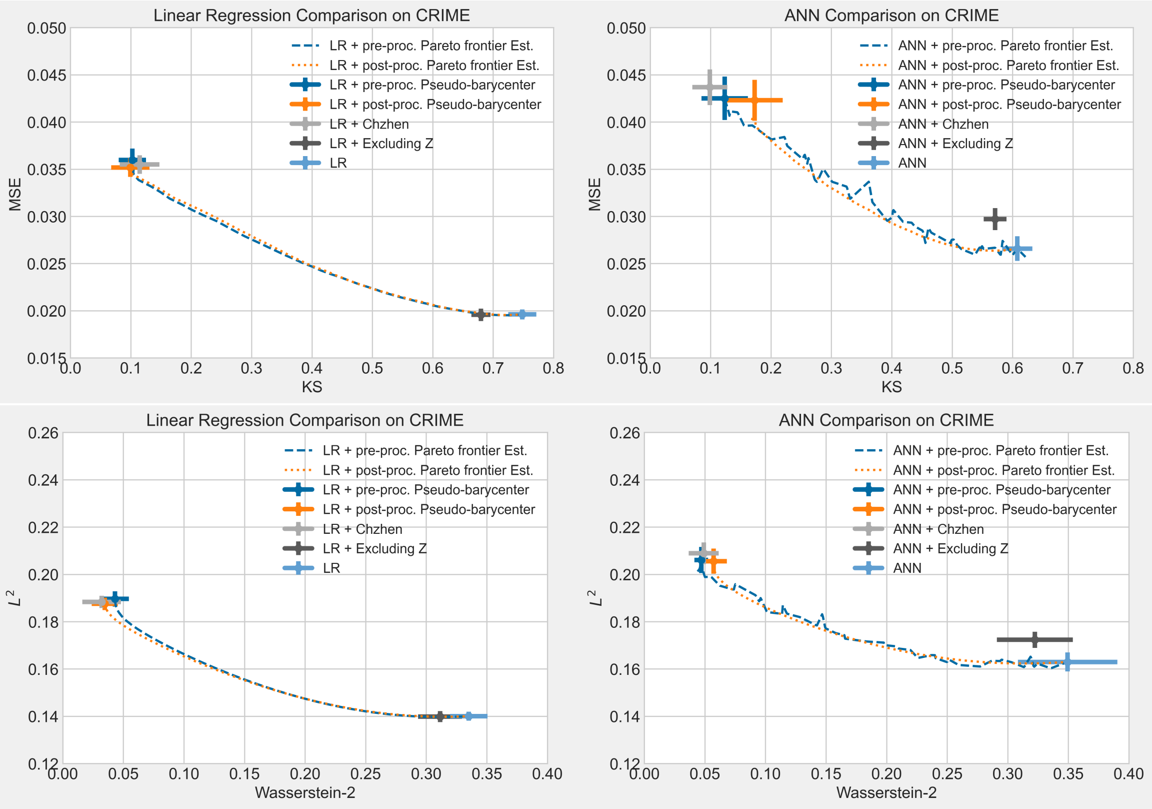


Figure 5: As shown above, the fair data representation method (+ pre-proc. Pareto frontier Est. or Pseudo-barycenter) achieved the same, if not better, performance as the exact barycenter method (+ Chzhen) in estimating the optimal learning outcome. In addition, the fair data representations method offers flexibility in choosing a desired (optimal) trade-off between utility and fairness.

For CRIME data, the small difference between the KS of learning outcome on the original data (LR and ANN) and the one on the data excluding the sensitive variable (LR and ANN + Excluding Z) implies a significant disparate impact. This observation and the multi-dimensional test below agree with the following statement in [17]: "Simply removing the 'protected attribute' is insufficient. As long as the model takes in features that are correlated with, say, gender or race, avoiding explicitly mentioning it will do little good."

In Figure 5, it is clear that the fair data representation methods (+ pre-proc. Pareto frontier Est. or Pseudo-barycenter) achieved the same, if not better, performance as the comparison method (+ Chzhen): the proposed method was outperformed by "+ Chzhen" with linear regression and outperformed "+ Chzhen" with artificial neural network, both by a narrow margin. But the performance of the fair data representation method is achieved at 4.735% of the time costs "+ Chzhen." In addition, the fair data representation method provides (an estimation of) the entire Pareto frontier and works for multivariate supervised learning (see Figure 6 below), whereas "+ Chzhen" only estimates the end point of the Pareto frontier and only works in the univariate learning.

Remark 7.1 *One possible explanation for the proposed method to outperform the exact post-processing Wasserstein barycenter method ("+ Chzhen") is the following: Although [18] is designed specifically for univariate learning and the KS distance by matching the sensitive marginal cumulative distribution functions, such matching on training data can lead to over-fitting. Therefore, the resulting optimal transport map fits the training data too well to be optimal for the test data.*

Next, we show the multivariate supervised learning on CRIME data to provide a high-dimensional baseline, to which later proposed machine learning fairness methods on high-dimensional data can compare. The vertical and horizontal axes are the L^2 test error and the $\mathcal{W}_2$ distance among sensitive groups. Hence, the more lower-left, the better the result.

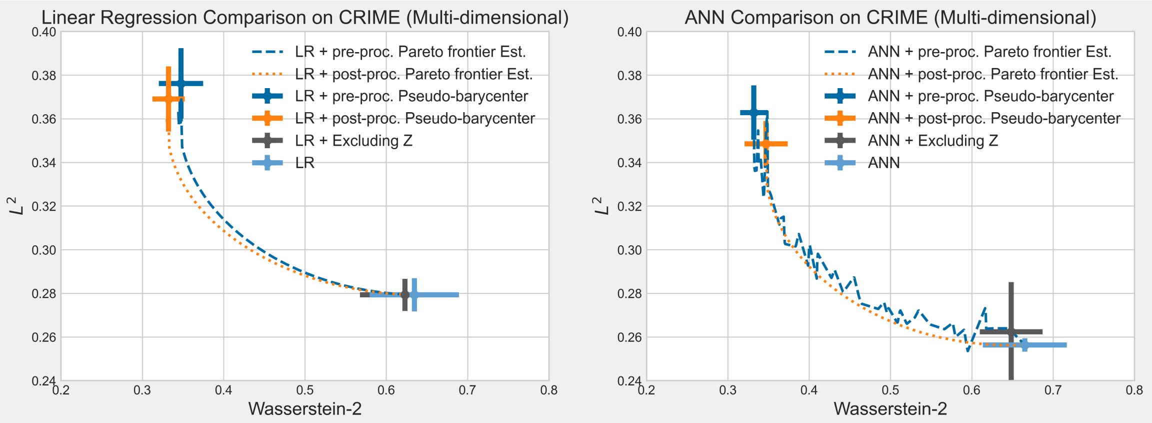
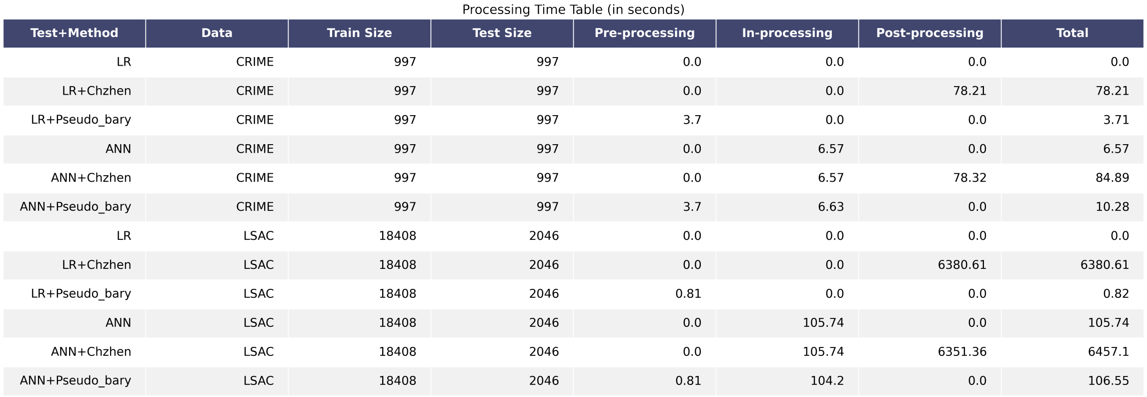


Figure 6: As shown above, the fair data representation method (+ pre-proc. Pareto frontier Est. or Pseudo-barycenter) achieves similar performance to the post-processing pseudo-barycenter method (+ post-proc. Pareto frontier Est. or Pseudo-barycenter).

Due to the relatively high dimensionality of X (87-dimensional) and Y (11-dimensional), the probabilistic dependence and correlation between the learning outcome and the sensitive variable Z becomes more difficult to remove. It is clear that (LR or ANN + Excluding Z) now removes almost none of the statistical disparity compared to the learning outcome on the original data.

To show the difference in practical computational cost among the comparison methods, we include the following processing time table, where the unit of time is second, and the simulations were run on a 2019 Macbook pro with Intel i9 processor.



Processing Time Table (in seconds)

Test+Method	Data	Train Size	Test Size	Pre-processing	In-processing	Post-processing	Total
LR	CRIME	997	997	0.0	0.0	0.0	0.0
LR+Chzhen	CRIME	997	997	0.0	0.0	78.21	78.21
LR+Pseudo_bary	CRIME	997	997	3.7	0.0	0.0	3.71
ANN	CRIME	997	997	0.0	6.57	0.0	6.57
ANN+Chzhen	CRIME	997	997	0.0	6.57	78.32	84.89
ANN+Pseudo_bary	CRIME	997	997	3.7	6.63	0.0	10.28
LR	LSAC	18408	2046	0.0	0.0	0.0	0.0
LR+Chzhen	LSAC	18408	2046	0.0	0.0	6380.61	6380.61
LR+Pseudo_bary	LSAC	18408	2046	0.81	0.0	0.0	0.82
ANN	LSAC	18408	2046	0.0	105.74	0.0	105.74
ANN+Chzhen	LSAC	18408	2046	0.0	105.74	6351.36	6457.1
ANN+Pseudo_bary	LSAC	18408	2046	0.81	104.2	0.0	106.55

Figure 7: As shown in the table above, the computational cost of the pseudo-barycenter method is significantly lower than the cost of the known post-processing methods: on average 7836 times faster on LSAC and 21 times faster on CRIME in a single train-test cycle for a single supervised learning model. Furthermore, in model selection or composition, the pre-processing time is a fixed one-time cost while the post-processing time is additive. (See point 4 below for a more detailed explanation)

Now, we show the major advantages of the proposed method compared to the post-processing ones, such as [18, 28, 24]:

1. Flexibility in Trade-off: The pre-processing method provides an estimation for the entire Pareto frontier and thereby allows practitioners to balance between prediction error and disparity. In contrast, the known post-processing method merely estimates the starting (left) point of the frontier.
2. Sensitive data privacy protection: The geodesics to the pseudo-barycenter allow practitioners to suppress the sensitive information remaining in the data to the desired level. That is, given the resulting suppressed data, anyone who has leaked data from the training or decision stage can merely extract the level of sensitive information up to the pre-determined remaining level. For example, if one chooses to suppress as much sensitive information as possible by setting $t = 1$, then it follows from the construction of dependent and independent pseudobarycenter, it is guaranteed that any unsupervised learning method that uses only the first two moments of the sample data distribution, such as the K-means and PCA, would be unable to extract any information about Z from $X^\dagger$ or $f_{Y^\dagger}(X^\dagger)$.
3. Computational efficiency in high-dimensional learning: As summarized in Figure 7, the computation of the pseudo-barycenter estimation of the optimal fair learning outcome is significantly faster than the computation of the exact barycenter via the post-processing matching cdf approach, especially on the LSAC data which has a larger sample size.
4. Flexibility in model selection, modification, and composition: in practice, one needs to repeat the training process multiple times to compare different supervised learning algorithms or parameters. The proposed fair data representation method has a fixed

pre-processing time while the processing time of post-processing methods is additive. For example, if a practitioner needs to compare linear regression and ANN on LSAC as shown in Figure 7 and repeat the training process N times for parameter tuning or validation purpose, the total processing time for pseudo-barycenter method is $0.81 + N(0.0025 + 104.2)$ while the processing time for the post-processing method is $N(0.003 + 6380.61 + 105.738 + 6351.36)$.

Acknowledgement

The authors want to thank the referees, whose profound and detailed feedback greatly enhanced the quality and clarity of this paper. The authors acknowledge support from NSF DMS-2027248, NSF DMS-2208356, NSF CCF-1934568, NIH R01HL16351, and DE-SC0023490.

A. Appendix: Proof of Results in Section 2

A.1 Proof of Lemma 2.1

Proof

$$\begin{aligned}
\mathcal{W}_2^2(\mu,\nu) &= \int ||x-y||^2 d\gamma^*(x,y) \\
&= \int ||((x-m_\mu)-(y-m_\nu)) + (m_\mu - m_\nu)||^2 d\gamma^*(x,y) \\
&= \int ||(x-m_\mu)-(y-m_\nu)||^2 d\gamma^*(x,y) + ||m_\mu - m_\nu||^2 \\
&\geq \mathcal{W}_2^2(\mu',\nu') + ||m_\mu - m_\nu||^2 \\
&= \int ||x-y||^2 d(\gamma')^*(x,y) + ||m_\mu - m_\nu||^2 \\
&= \int ||(x+m_\mu)-(y+m_\nu)||^2 d(\gamma')^*(x,y) \\
&\geq \mathcal{W}_2^2(\mu,\nu)
\end{aligned}$$

where γ^* and $(\gamma')^*$ denote the optimal transport plan for (μ,ν) and (μ',ν'), respectively. The first inequality results from the fact that $\gamma'(x,y) := \gamma^*(x-m_\mu, y-m_\nu) \in \prod(\mu',\nu')$, the second inequality from $\gamma(x,y) := (\gamma')^*(x+m_\mu, y+m_\nu) \in \prod(\mu,\nu)$, and the equalities from direct expansion. ■

A.2 Proof of Lemma 2.3

Proof Existence and uniqueness follow directly from Theorem 2.1. For the equivalent multi-marginal coupling problem, there exists an optimal solution $\gamma^* = \mathcal{L}(\{X_z\}_z)$. It follows from Remark 2.3 that $\bar{X} = T(\{X_z\}_z)$ where $\mathcal{L}(\bar{X})$ is the Wasserstein barycenter. Therefore, the Gaussianity of barycenter results from linearity of T in the finite $|\mathcal{Z}|$ case, and the fact that the set of Gaussian distribution is closed in $(\mathcal{P}_{2,ac}, \mathcal{W}_2)$ when $|\mathcal{Z}|$ is infinite. The characterization equation is proved in the case of finite $|\mathcal{Z}|$ in [2]. For infinite $|\mathcal{Z}|$, the equation still holds due to the continuity of the covariance function on $(\mathcal{P}_{2,ac}, \mathcal{W}_2)$. The sufficiency and necessity of the equation follows from the following characterization of the barycenter via Brenier's maps $\{T_{\bar{X}X_z}\}_z$ derived in [2]:

$$\int_{\mathcal{Z}} T_{\bar{X}X_z} d\lambda(z) = Id. \tag{96}$$

It follows from the explicit form of $\{T_{\bar{X}X_z}\}_z$ in Lemma 2.2 that

$$\begin{aligned}
&\int_{\mathcal{Z}} T_{\bar{X}X_z} d\lambda(z) = \int_{\mathcal{Z}} \Sigma_{\bar{X}}^{-\frac{1}{2}} (\Sigma_{\bar{X}}^{\frac{1}{2}} \Sigma_{X_z} \Sigma_{\bar{X}}^{\frac{1}{2}})^{\frac{1}{2}} \Sigma_{\bar{X}}^{-\frac{1}{2}} d\lambda(z) = Id \\
\Longleftrightarrow\ & \Sigma_{\bar{X}}^{\frac{1}{2}} \Sigma_{\bar{X}}^{-\frac{1}{2}} \int_{\mathcal{Z}} (\Sigma_{\bar{X}}^{\frac{1}{2}} \Sigma_{X_z} \Sigma_{\bar{X}}^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) \Sigma_{\bar{X}}^{-\frac{1}{2}} \Sigma_{\bar{X}}^{\frac{1}{2}} = \Sigma_{\bar{X}}^{\frac{1}{2}} Id \Sigma_{\bar{X}}^{\frac{1}{2}} \\
\Longleftrightarrow\ & \int_{\mathcal{Z}} (\Sigma_{\bar{X}}^{\frac{1}{2}} \Sigma_{X_z} \Sigma_{\bar{X}}^{\frac{1}{2}})^{\frac{1}{2}} d\lambda(z) = \Sigma_{\bar{X}}.
\end{aligned}$$

■

B. Appendix: Proof of Results in Section 4

B.1 Proof of Lemma 4.1

Proof First, it follows from the triangle inequality that

$$\mathcal{W}_2(\mu_0, \mu_1) \leq \mathcal{W}_2(\mu_0, \mu_s) + \mathcal{W}_2(\mu_s, \mu_t) + \mathcal{W}_2(\mu_t, \mu_1)$$

for any $s, t \in [0, 1]$. On the other hand, it follows from the definition of μ_t that for $s, t \in [0, 1]$

$$\begin{aligned}
\mathcal{W}_2^2(\mu_s, \mu_t) &\leq \int_{(\mathbb{R}^d)^2} ||x - y||^2 d(\pi_s)_\sharp \gamma(x) \otimes d(\pi_t)_\sharp \gamma(y) \\
&= \int_{(\mathbb{R}^d)^2} ||\pi_s(x, y) - \pi_t(x, y)||^2 d\gamma(x, y) \\
&= \int_{(\mathbb{R}^d)^2} ||(1 - s)x + sy - (1 - t)x - ty||^2 d\gamma(x, y) \\
&= \int_{(\mathbb{R}^d)^2} ||(t - s)x - (t - s)y||^2 d\gamma(x, y) \\
&= |t - s|^2 \int_{(\mathbb{R}^d)^2} ||x - y||^2 d\gamma(x, y) = |t - s|^2 \mathcal{W}_2^2(\mu_0, \mu_1),
\end{aligned}$$

where the first equation results from definition of $\mathcal{W}_2$. Given the above two facts, we complete the proof by contradiction. Assume $\exists s, t \in [0, 1]$ such that $\mathcal{W}_2(\mu_s, \mu_t) < |t - s|\mathcal{W}_2(\mu_0, \mu_1)$, then

$$\begin{aligned}
\mathcal{W}_2(\mu_0, \mu_1) &\leq \mathcal{W}_2(\mu_0, \mu_s) + \mathcal{W}_2(\mu_s, \mu_t) + \mathcal{W}_2(\mu_t, \mu_1) \\
&< |s|\mathcal{W}_2(\mu_0, \mu_1) + |t - s|\mathcal{W}_2(\mu_0, \mu_1) + |1 - t|\mathcal{W}_2(\mu_t, \mu_1) \\
&= \mathcal{W}_2(\mu_0, \mu_1).
\end{aligned}$$

■

B.2 Proof of Theorem 4.1

Proof First, we derive the inequality from the triangle inequality and the optimality of $\{T(\cdot, z)\}_z$: Let $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ be an arbitrary measurable function. It follows that

$$\begin{aligned}
V &\leq (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - \overline{f(X,Z)}_z||_2^2 d\lambda(z))^{\frac{1}{2}} \\
&\leq L(f(X,Z)) + (\int_{\mathcal{Z}} ||f(X,Z)_z - \overline{f(X,Z)}_z||_2^2 d\lambda(z))^{\frac{1}{2}} \\
&\leq L(f(X,Z)) + (\int_{\mathcal{Z}} \mathcal{W}_2^2(\mathcal{L}(f(X,Z)_z), \overline{\mathcal{L}(f(X,Z)_z)}) d\lambda(z))^{\frac{1}{2}} \\
&= L(f(X,Z)) + (\frac{1}{2}\int_{\mathcal{Z}^2} \mathcal{W}_2^2(\mathcal{L}(f(X,Z)_{z_1}), \mathcal{L}(f(X,Z)_{z_2})) d\lambda(z_1) d\lambda(z_2))^{\frac{1}{2}} \\
&= L(f(X,Z)) + \frac{1}{\sqrt{2}} D(f(X,Z)).
\end{aligned}$$

Here, the penultimate equation results from the fact that, for any $\{\nu_z\}_z \subset \mathcal{P}_{2,ac}(\mathbb{R}^d)$,

$$\int_{\mathcal{Z}^2} \mathcal{W}_2^2(\nu_{z_1}, \nu_{z_2}) d\lambda(z_1) d\lambda(z_2) = 2\int_{\mathcal{Z}} \mathcal{W}_2^2(\nu_z, \bar{\nu}) d\lambda(z), \tag{97}$$

where $\bar{\nu}$ is the Wasserstein barycenter of $\{\nu_z\}_z$. Now, we show that the lower bound is achieved if and only if $f(X,Z) = T(t)(\mathbb{E}(Y|X,Z), Z), t \in [0,1]$. Let $t \in [0,1]$, $T_z := T(\cdot, z)$, and $\mu_z := \mathcal{L}(\mathbb{E}(Y|X,Z)_z)$. It follows from Lemma 4.1 and Remark 4.1 that:

$$\begin{aligned}
V &= (\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda(z))^{\frac{1}{2}} \\
&\leq (\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, T_z(t)_\sharp \mu_z) d\lambda(z))^{\frac{1}{2}} + (\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp \mu_z, \bar{\mu}) d\lambda(z))^{\frac{1}{2}} \\
&= (t^2 \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda(z))^{\frac{1}{2}} + ((1-t)^2 \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda(z))^{\frac{1}{2}} \\
&= tV + (1-t)V = V.
\end{aligned}$$

Therefore, the second inequality is an equality where the first term is $L(T(t))$:

$$\begin{aligned}
L(T(t)) &= (\int_{\mathcal{Z}} ||\mathbb{E}(Y|X,Z)_z - T_z(t)(\mathbb{E}(Y|X,Z)_z)||_2^2 d\lambda(z))^{\frac{1}{2}} \\
&= (\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, T_z(t)_\sharp \mu_z) d\lambda(z))^{\frac{1}{2}} \\
&= t(\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda(z))^{\frac{1}{2}} = tV.
\end{aligned}$$

For the second term, we claim that it equals $\frac{1}{\sqrt{2}}D(T(t))$. To see this, we need to first show $\overline{T_z(t)_\sharp\mu_z} = \bar{\mu}$. Indeed, if not, then $\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \overline{T_z(t)_\sharp\mu_z})d\lambda(z)$ is strictly less than $\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \bar{\mu})d\lambda(z)$ by the definition and uniqueness of $\overline{T_z(t)_\sharp\mu_z}$. It follows that

$$\begin{aligned}
&(\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \overline{T_z(t)_\sharp\mu_z})d\lambda(z))^{\frac{1}{2}} \\
\leq&(\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, T_z(t)_\sharp\mu_z)d\lambda(z))^{\frac{1}{2}} + (\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \overline{T_z(t)_\sharp\mu_z})d\lambda(z))^{\frac{1}{2}} \\
<&L(T(t)) + (\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \bar{\mu})d\lambda(z))^{\frac{1}{2}} \\
=&(\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu})d\lambda(z))^{\frac{1}{2}},
\end{aligned}$$

which contradicts the definition and uniqueness of $\bar{\mu}$. Therefore,

$$\begin{aligned}
D(T(t)) &= (\int_{\mathcal{Z}^2} \mathcal{W}_2^2(T_{z_1}(t)_\sharp\mu_{z_1}, T_{z_2}(t)_\sharp\mu_{z_2})d\lambda(z_1)d\lambda(z_2))^{\frac{1}{2}} \\
&= (2\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \overline{T_z(t)_\sharp\mu_z})d\lambda(z))^{\frac{1}{2}} \\
&= \sqrt{2}(\int_{\mathcal{Z}} \mathcal{W}_2^2(T_z(t)_\sharp\mu_z, \bar{\mu})d\lambda(z))^{\frac{1}{2}} \\
&= \sqrt{2}((1-t)^2\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu})d\lambda(z))^{\frac{1}{2}} \\
&= \sqrt{2}(1-t)V.
\end{aligned}$$

That completes the proof. ■

C. Appendix: Proof of Results in Section 5

C.1 Proof of $\tilde{X} \perp Z$ implies $\mathbb{E}(Y_z|\tilde{X}) = \mathbb{E}(Y|\tilde{X}, Z)_z$

Proof Let $\tilde{X} \perp Z$ and assume for contradiction that $\mathbb{E}(Y_z|\tilde{X}) \neq \mathbb{E}(Y|\tilde{X}, Z)_z$. Then, we have

$$\begin{aligned}
||Y - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 &= \int_{\mathcal{Z}} ||Y_z - f^*(\tilde{X}, Z)_z||_2^2 d\lambda \\
&= \int_{\mathcal{Z}} ||Y_z - f^*(\tilde{X}, z)||_2^2 d\lambda \\
&> \int_{\mathcal{Z}} ||Y_z - \mathbb{E}(Y_z|\tilde{X})||_2^2 d\lambda \\
&= \int_{\mathcal{Z}} ||Y_z - \tilde{f}_z(\tilde{X})||_2^2 d\lambda
\end{aligned}$$

where the first line follows from disintegration and the fact that there exists a measurable function $f^* : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ such that $f^*(\tilde{X}, Z) = \mathbb{E}(Y|\tilde{X}, Z)$, the second from $\tilde{X} \perp Z$, the third line follows from orthogonal projection property of conditional expectation and the assumption, and the forth from the fact that there exists a measurable function $\tilde{f}_z : \mathcal{X} \to \mathcal{Y}$ such that $\tilde{f}_z(\tilde{X}) = \mathbb{E}(Y_z|\tilde{X})$. Now, define $\tilde{f} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ by $\tilde{f}(\cdot, z) := \tilde{f}_z$ for λ-a.e. $z \in \mathcal{Z}$. It follows that

$$\begin{aligned}
||Y - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 &> \int_{\mathcal{Z}} ||Y_z - \tilde{f}_z(\tilde{X})||_2^2 d\lambda \\
&= \int_{\mathcal{Z}} ||Y_z - \tilde{f}(\tilde{X}, z)||_2^2 d\lambda \\
&= ||Y - \tilde{f}(\tilde{X}, Z)||_2^2 \\
&= ||Y - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 + ||\mathbb{E}(Y|\tilde{X}, Z) - \tilde{f}(\tilde{X}, Z)||_2^2.
\end{aligned}$$

That implies $||\mathbb{E}(Y|\tilde{X}, Z) - \tilde{f}(\tilde{X}, Z)||_2^2 < 0$, a contradiction. This completes the proof. ■

C.2 Proof of Lemma 5.1

Proof Let $\tilde{X}, \tilde{X}' \in \{\tilde{X} \in \mathcal{D}_{\mathcal{X}} : \tilde{X} \perp Z\}$ satisfy $\sigma(\tilde{X}') \subset \sigma(\tilde{X})$. We have

$$\begin{aligned}
&||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}|\tilde{X}, Z)||_2^2 - ||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}'|\tilde{X}', Z)||_2^2 \\
=&||\mathbb{E}(Y|X, Z) - \overline{\mathbb{E}(Y|\tilde{X}, Z)}||_2^2 - ||\mathbb{E}(Y|X, Z) - \overline{\mathbb{E}(Y|\tilde{X}', Z)}||_2^2
\end{aligned}$$

Notice that

$$||\mathbb{E}(Y|X, Z) - \overline{\mathbb{E}(Y|\tilde{X}, Z)}||_2^2 = ||\mathbb{E}(Y|X, Z) - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda$$

where $\mu_z := \mathcal{L}(\mathbb{E}(Y|\tilde{X}, Z)_z)$ and $\bar{\mu} := \overline{\mathcal{L}(\mathbb{E}(Y|\tilde{X}, Z))}$. Also, we define μ'_z and $\bar{\mu}'$ analogously to have

$$\begin{aligned}
&||\mathbb{E}(Y|X, Z) - \overline{\mathbb{E}(Y|\tilde{X}', Z)}||_2^2 \\
=&||\mathbb{E}(Y|X, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \bar{\mu}') d\lambda \\
=&||\mathbb{E}(Y|X, Z) - \mathbb{E}(Y|\tilde{X}, Z)||_2^2 + ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \bar{\mu}') d\lambda.
\end{aligned}$$

Combining the above, we have

$$\begin{aligned}
&||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}|\tilde{X}, Z)||_2^2 - ||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}'|\tilde{X}', Z)||_2^2 \\
=&\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda - \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \bar{\mu}') d\lambda - ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2.
\end{aligned}$$

It remains to show that $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda < \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \bar{\mu}') d\lambda + ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2$. Indeed, assume for contradiction that $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu'_z, \bar{\mu}') d\lambda + ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2 \leq$

$\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda$, then we have

$$\begin{aligned}\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}') d\lambda &\leq ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2 + \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z', \bar{\mu}') d\lambda \\ &\leq \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda.\end{aligned}$$

This contradicts the optimality and uniqueness of $\bar{\mu}$ by Lemma 3.1. Therefore, we prove by contradiction that $\int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \bar{\mu}) d\lambda < \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z', \bar{\mu}') d\lambda + ||\mathbb{E}(Y|\tilde{X}, Z) - \mathbb{E}(Y|\tilde{X}', Z)||_2^2$ and, hence,

$$||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}|\tilde{X}, Z)||_2^2 - ||\mathbb{E}(Y|X, Z) - \mathbb{E}(\bar{Y}'|\tilde{X}', Z)||_2^2 < 0.$$

That completes the proof. ■

C.3 Proof of Lemma 5.2

Proof We first prove $\sigma((\bar{X}, Z)) = \sigma((X, Z))$. Since $\mathcal{L}(X_z) \subset \mathcal{P}_{2,ac}$, it follows from Lemma 3.1 that there exists a measurable map $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ such that $T(X_z, z) = \bar{X}_z$ λ-a.e., where $\bar{X}$ denotes the Wasserstein barycenter of $\{X_z\}_z$. Define $T \otimes Id|_{\mathcal{Z}} : \mathcal{X} \times \mathcal{Z} \to \mathcal{X} \times \mathcal{Z}$, we have $T \otimes Id|_{\mathcal{Z}}$ is $\mathcal{X} \times \mathcal{Z}/\mathcal{X} \times \mathcal{Z}$-measurable and satisfies $T \otimes Id|_{\mathcal{Z}}((X, Z)) = (\bar{X}, Z)$. That implies $\sigma((\bar{X}, Z)) \subset \sigma((X, Z))$. Furthermore, since $\mathcal{L}(\bar{X}) \in \mathcal{P}_{2,ac}$, it follows from Brenier's theorem [11] that there exists $T^{-1}(\cdot, z)$ such that $T^{-1}(\bar{X}_z, z) = X_z$. Therefore, we have $(T \otimes Id|_{\mathcal{Z}})^{-1} = T^{-1} \otimes Id|_{\mathcal{Z}}$ is $\mathcal{X} \times \mathcal{Z}/\mathcal{X} \times \mathcal{Z}$-measurable and satisfies $(T \otimes Id|_{\mathcal{Z}})^{-1}((\bar{X}, Z)) = (X, Z)$. That implies $\sigma((X, Z)) \subset \sigma((\bar{X}, Z))$. That completes the proof of $\sigma((\bar{X}, Z)) = \sigma((X, Z))$. Now, we show $\sigma(\tilde{X}) \subset \sigma(\bar{X})$. From the construction of $\tilde{X}$, we have $\sigma((\tilde{X}, Z)) \subset \sigma((\bar{X}, Z)) = \sigma((X, Z))$. But $\tilde{X} \perp Z$ implies that, for any $B_X \in \mathcal{B}_{\mathcal{X}}$, we can construct $B_X \times \mathcal{Z} \in \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}$. In addition, due to $\sigma((\tilde{X}, Z)) \subset \sigma((\bar{X}, Z))$, there exists $B'_{XZ} \in \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Z}}$ such that $(\bar{X}, Z)^{-1}(B'_{XZ}) = (X, Z)^{-1}(B_X \times \mathcal{Z})$. Lastly, $\bar{X} \perp Z$ also implies that there exists $B'_X \in \mathcal{B}_{\mathcal{X}}$ satisfying $B'_{XZ} = B'_X \times \mathcal{Z}$. It follows that

$$\tilde{X}^{-1}(B_X) = (\tilde{X}, Z)^{-1}(B_X \times \mathcal{Z}) = (X, Z)^{-1}(B'_X \times \mathcal{Z}) = X^{-1}(B'_X) \tag{98}$$

Since our choice of $B_X \in \mathcal{B}_{\mathcal{X}}$ is arbitrary, it follows that $\sigma(\tilde{X}) \subset \sigma(\bar{X})$. Finally, since our choice of $\tilde{X} \in \{\tilde{X} \in \mathcal{D}|_{\mathcal{X}} : \tilde{X} \perp Z\}$ is arbitrary, we are done. ■

References

[1] B. L. Adamson. Ricci v. DeStefano: Procedural Activism (?). *National Black Law Journal (University of California, Los Angeles)*, 24:11–01, 2011.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] J. M. Altschuler and E. Boix-Adsera. Wasserstein barycenters are NP-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

[4] P. C. Álvarez-Esteban, E. Del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[6] J. B. Aristotle et al. *The complete works of Aristotle*, volume 2. Princeton University Press Princeton, 1984.

[7] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[8] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[9] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2009.

[10] A. W. Blumrosen. Strangers in paradise: Griggs v. Duke Power Co. and the concept of employment discrimination. *Mich. L. Rev.*, 71:59, 1972.

[11] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[12] T. Calders and I. Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 43–57. Springer, 2013.

[13] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[14] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

[15] G. Carlier and I. Ekeland. Matching for teams. *Economic theory*, 42:397–418, 2010.

[16] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[17] B. Christian. *The alignment problem: Machine learning and human values.* WW Norton & Company, 2020.

[18] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:7321–7331, 2020.

[19] J. M. Cooper, D. S. Hutchinson, et al. *Plato: complete works.* Hackett Publishing, 1997.

[20] J. A. Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Diaz. On lower bounds for the l2-Wasserstein metric in a Hilbert space. *Journal of Theoretical Probability*, 9(2):263–283, 1996.

[21] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[22] I. Ekeland. Existence, uniqueness and efficiency of equilibrium in hedonic markets with multidimensional types. *Economic Theory*, 42:275–315, 2010.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[24] T. L. Gouic, J.-M. Loubes, and P. Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020.

[25] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25 (7):1445–1459, 2012.

[26] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[27] L. Hu and Y. Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[28] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR, 2020.

[29] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[30] Y.-H. Kim and B. Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017.

[31] T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probability Theory and Related Fields*, 168:901–917, 2017.

[32] R. J. McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[33] E. O. of the President. Big data: Seizing opportunities, preserving values. *President PACT report*, 2014.

[34] B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[35] M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

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

[37] C. Silvia, J. Ray, S. Tom, P. Aldo, J. Heinrich, and A. John. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), pages 3633–3640, 2020.

[38] L. Sweeney. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, 2013.

[39] E. G. Tabak and G. Trigila. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics*, 71(1):163–199, 2018.

[40] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[41] C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[42] L. F. Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998.

[43] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[44] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[45] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, J. Chen, and A. Sudjianto. Bias, Fairness, and Accountability with AI and ML Algorithms. *arXiv preprint arXiv:2105.06558*, 2021.