Geometry of Relaxed Fair Regression: A Unified Framework for Aware and Unaware Settings

Marie Generali Lince
Inria, Fairplay joint team
CREST, ENSAE, IP Paris
Palaiseau, France
marie.generali@ensae.fr

Vincent Divol
CREST, ENSAE, IP Paris
Palaiseau, France
vincent.divol@ensae.fr

Remi Flamary
CMAP, Ecole Polytechnique
Palaiseau, France
remi.flamary@polytechnique.edu

Solenne Gaucher
CMAP, Ecole Polytechnique
Palaiseau, France
solenne.gaucher@polytechnique.edu

Patrick Loiseau
Inria, Fairplay joint team
Palaiseau, France
patrick.loiseau@ensae.fr

Abstract

Fairness-accuracy trade-offs are a central concern in the deployment of fairness-aware machine learning methods. When sensitive attributes are unavailable at inference time–the so called unawareness setting, principled methods for obtaining accurate predictions under relaxed fairness constraints are largely missing. In this work, we address this gap by formulating regression under a demographic parity penalty as an optimal transport problem. Our framework, unifies both the *aware* and *unaware* settings and characterizes optimal prediction functions via optimal transport maps, under both squared Wasserstein-2 and Total Variation penalties. These results reveal that the choice of penalty reflects fundamentally different fairness philosophies: the Wasserstein penalty induces a smooth, population-wide compromise, while Total Variation enforces parity for a subset of individuals. Building on these theoretical characterizations, we propose FAUST, an algorithm that is simple to implement, computationally efficient, and consistently matches or outperforms state-of-the-art baselines on real-world benchmarks.

1 Introduction

As machine learning models are increasingly deployed in high-stakes domains, mitigating algorithmic bias has become a critical objective. A significant portion of the fairness literature addresses group fairness, imposing fair distribution of the prediction across groups. Among these, Demographic Parity (DP) has received considerable attention due to its simplicity and interpretability, aiming to ensure that predictions are statistically independent of the sensitive attribute. Mitigation strategies typically intervene on the training data (*pre-processing*), the learning algorithm (*in-processing*), or the model's predictions (*post-processing*) [6, 17]. Post-processing methods are particularly attractive as they allow for the retrofitting of any black-box predictor without requiring retraining, making them applicable in real-world settings where the base model is already deployed. In this work, we focus on post-processing for regression under the DP constraint.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The strategy for enforcing DP depends on data availability: in the *aware* setting, the sensitive attribute is observed alongside standard features, while in the *unaware* setting, it is unavailable at inference time. A classical heuristic to circumvent this constraint is to explicitly estimate the sensitive attribute and apply group-aware corrections using this plug-in estimator, but **plug-in approaches generally fail to achieve fairness** (see Fig. 1 Section C). Furthermore, enforcing *exact* DP in either setting is often too rigid, causing severe degradation in predictive utility. This limitation motivates the study of *relaxed* fairness, which allows for controlled violations to optimize the accuracy-fairness trade-off.

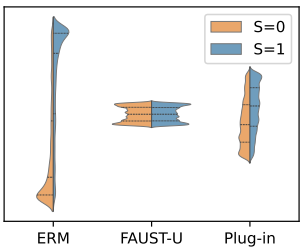

Figure 1: **FAUST-U vs. Plug-In.** Marginal distributions of the predictions in each group (on data from Section 2.4). The Plug-in approach fails retaining the ERM gap, whereas FAUST-U successfully aligns group distributions.

Related works. Empirical Risk Minimization (ERM) under fairness penalties or relaxed fairness constraints is common within the *in-processing* paradigm. Numerous methods directly incorporate fairness penalties or bounded constraints into the learning objective during model training [1, 7, 26, 29, 32, 41, 13]. More recently, optimal transport (OT) has been leveraged within this paradigm to regularize latent representations and enforce relaxed fairness even when sensitive attributes are unobserved [30]. These approaches inherit a core limitation of in-processing methods: they presuppose control over the algorithm's training procedure, which is unrealistic in many real-world settings. By contrast, OT provides a natural geometric framework for enforcing DP through *post-processing* approaches. In the *aware* setting, exact fairness is achieved by mapping group marginal distributions of the Bayes predictor to their Wasserstein barycenter [12, 10, 24, 23, 15, 7, 32], and closed-form relaxations exist to control the trade-off between fairness and accuracy [11]. In the *unaware* setting, achieving parity often relies on noisy proxies or robust optimization techniques [9, 4]. Recent work has characterized the optimal regression function under DP using OT over 2D measures [16]. Yet extending the fairness relaxations to the unaware setting remains an open challenge. State-of-the-art methods currently bypass characterization of the optimal fair predictor: FairReg [38] relies on grid-based discretization and stochastic optimization, while the recent FRAPPÉ framework [14] learns an additive neural network correction via stochastic gradient descent. This highlights a significant gap in the literature: while exact fairness under unawareness admits an OT formulation, its relaxed counterpart is solved via grid-based approximations or gradient-based methods.

Outline and contributions. In this paper, we bridge this gap by establishing pointwise closed-form reductions for relaxed fairness across both *aware* and *unaware* settings, focusing on the two-groups case. While restrictive, the two-groups assumption is common in the literature [33] and captures the practically important scenario in which one group corresponds to a protected class, while already revealing key phenomena. Our main contributions are as follows:

- **A unified framework for relaxed fair regression:** In Section 2.2, we formulate relaxed fair regression as a risk minimization problem penalized by a fairness discrepancy metric as
$$f^*_{\lambda,\mathcal{D}} \in \arg\min_f \ \mathbb{E}\left[(f(X) - \eta(X))^2\right] + \lambda\, \mathcal{U}_{\mathcal{D}}(f),$$
accommodating different fairness penalties ($\mathcal{W}_2$ and TV) across both aware and unaware settings.
- **Closed-form solutions via OT reduction:** In Section 2.3 we show that, the penalized relaxation reduces to a standard OT problem. Unlike existing state-of-the-art methods, **we characterize the optimal predictor** for almost-fair prediction functions under both $\mathcal{W}_2^2$ and TV penalties and provide closed-form solution via OT maps.
- **Practical implications:** In Section 2.4, we make explicit the trade-offs inherent in the choice of fairness penalty and contrast their practical implications. We show that the relaxation penalty is not just a numerical convenience, but reflects fundamentally different approaches to fairness.
- **Implementation and experiments:** In Section 3, we demonstrate how these closed-form solutions can be efficiently estimated from finite samples to build out-of-sample fair predictors. In Section 4, our theory-driven approach is empirically validated on the Law School and Communities and Crime datasets. [1].

[1] Our code is available at `https://github.com/mgeneralince/FAUST`.

Notation. Let $\mathcal{P}(\Omega)$ be the set of probability measures on Ω, and $\mathcal{L}(\cdot)$ the law of a random variable. For $\mu \in \mathcal{P}(\mathbb{R})$, denote its cumulative distribution function (c.d.f.) by F_μ. For a measurable map T, $T_\# \mu$ is the pushforward measure. The set of couplings with marginals μ and ν is denoted $\Pi(\mu, \nu)$. The OT cost between μ and ν under a cost c is defined by $\mathrm{OT}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y)\, \mathrm{d}\pi(x, y)$. $\mathcal{W}_2^2(\mu, \nu)$ is the squared Wasserstein-2 distance, i.e., $\mathrm{OT}_{c_2}(\mu, \nu)$ with cost $c_2(x, y) = (x - y)^2$, and $\mathrm{TV}(\mu, \nu)$ the Total Variation distance, i.e. $\mathrm{OT}_{c_0}(\mu, \nu)$ with cost $c_0(x, y) = \mathbb{1}_{x \neq y}$ [31, Proposition 4.7]. Finally, for $a \in \mathbb{R}$, $a_+ = \max(a, 0)$ and $a_- = -\min(a, 0)$.

2 Unified framework for aware and unaware relaxed fair regression

We consider a regression problem where individuals are characterized by a non-sensitive feature X taking values in some set $\mathcal{X} \subset \mathbb{R}^d$, a binary sensitive attribute $S \in \{+, -\}$ with prior probabilities $p^+ = \mathbb{P}(S = +)$ (resp. $p^- = \mathbb{P}(S = -)$), and a continuous target variable $Y \in \mathbb{R}$ with $\mathbb{E}[Y^2] < \infty$. The law of X in group $S = +$ (resp. in group $S = +$) is denoted $\chi^+ = \mathcal{L}(X|S = +)$ (resp. $\chi^- = \mathcal{L}(X|S = -)$). Abusing notations, let $\eta(x, s) = \mathbb{E}[Y \mid X = x, S = s]$ and $\eta(x) = \mathbb{E}[Y \mid X = x]$.

In regression problems, the goal is to predict an outcome $Y \in \mathbb{R}$ from either the pair (X, S) (in the awareness setting) or from X (in the unawareness setting). Then, under mean square error, the Bayes optimal predictor is given by the conditional expectation η. Unfortunately, its predictions often reflect systemic biases toward S [6]. Among the criteria characterizing fair prediction, DP is satisfied by a prediction function f if its predictions are statistically independent of S: $(f(X, S) \perp\!\!\!\perp S)$ in the awareness setting, and $(f(X) \perp\!\!\!\perp S)$ in the unawareness setting.

2.1 Existing results for exact fairness

Exact fairness under awareness. In the awareness setting, Chzhen et al. [12] show that finding the optimal fair predictor is equivalent to solving the following Wasserstein barycenter problem. For $s \in \{+, -\}$, let $\mu_\eta^s = \eta(\cdot, s)_\# \chi^s$. When the distributions μ_η^+ and μ_η^- are non atomic, the authors show that the distribution of the optimal fair prediction ν^* corresponds to the Wasserstein barycenter

$$\nu^* = \arg\min_{\nu \in \mathcal{P}(\mathbb{R})} \left\{ p^+ \mathcal{W}_2^2(\mu_\eta^+, \nu) + p^- \mathcal{W}_2^2(\mu_\eta^-, \nu) \right\}. \tag{1}$$

Since μ_η^+ and μ_η^- are one-dimensional, the barycenter problem admits a closed-form OT map. The optimal fair prediction $f^*(x, s)$ is obtained by applying this map (transporting from μ_η^s to the barycenter ν^*) to $\eta(x, s)$. We detail the closed-form solution in Section B.

Exact fairness under unawareness. In the *unaware* setting, S is unobserved at inference time. As established by Divol and Gaucher [16], the optimal fair prediction implicitly infers S by basing its predictions on the signed group probability Δ given by

$$\Delta(x) = \frac{\mathbb{P}(S = + \mid X = x)}{p^+} - \frac{\mathbb{P}(S = - \mid X = x)}{p^-}. \tag{2}$$

Let us denote $\left((\chi^+ - \chi^-)_+, (\chi^+ - \chi^-)_-\right)$ the normalized Jordan decomposition of the signed measure $\chi^+ - \chi^-$, where the normalization constant is chosen so that $(\chi^+ - \chi^-)_+$ and $(\chi^+ - \chi^-)_-$ are probability measures [2]. Then, the fair regression problem reduces to an OT problem between the probability measures $\mu_{\eta,\Delta}^+$ and $\mu_{\eta,\Delta}^-$ on $\mathbb{R} \times \mathbb{R}$, where

$$\mu_{\eta,\Delta}^+ = (\eta(\cdot), \Delta(\cdot))_\# \left(\chi^+ - \chi^-\right)_+ \qquad \text{and} \qquad \mu_{\eta,\Delta}^- = (\eta(\cdot), \Delta(\cdot))_\# \left(\chi^+ - \chi^-\right)_-.$$

By construction, the probability measure $\mu_{\eta,\Delta}^+$ (resp. $\mu_{\eta,\Delta}^-$) only gives weight to the half space $\{(h, d) \in \mathbb{R}^2 : d > 0\}$ (resp. the half space $\{(h, d) \in \mathbb{R}^2 : d < 0\}$). Then, the fair regression problem can be reduced to an OT problem under the following assumption.

Assumption 1. The measures $\mu_{\eta,\Delta}^+$ and $\mu_{\eta,\Delta}^-$ give zero mass to graphs of functions in the sense that for any measurable function $F : \mathbb{R}^* \to \mathbb{R}$ and $s \in \{+, -\}$, $\mu_{\eta,\Delta}^s(\{(F(d), d) : d \neq 0\}) = 0$.

Under Assumption 1, Divol and Gaucher [16] shows that finding the optimal fair predictor reduces to solving the following barycenter problem

[2] See Divol and Gaucher [16], Section 3.2.1 or Section A.1 for a construction of these measures.

$$\inf_{\nu \in \mathcal{P}(\mathbb{R})} \left\{ \mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \nu) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \nu) \right\}, \tag{P_∞}$$

with cost $c_u(z, y) = \frac{|h-y|^2}{|d|}$ specific to the unaware geometry, where $z = (h, d) \in \mathbb{R} \times \mathbb{R}$. Note that this is a two-to-one dimensional OT problem, which in general does not admit a closed-form solution but can be solved numerically [37].

Reducing awareness to unawareness. As noted by Divol and Gaucher [16], the results for the awareness setting can be recovered from those in the unawareness setting by assuming that the sensitive attribute S is included in the features X, or equivalently that S is X-measurable. In this case, $\Delta(x) = \frac{s}{p^s}$. Then, Assumption 1 reduces to assuming non-atomicity of the measures μ^s_η, a condition that is necessary for the results of Chzhen et al. [12] to hold. Similarly, the transport cost simplifies as $c((\eta(x), \Delta(x)), y) = p^s(\eta(x) - y)^2$. Consequently, the OT problem defined in the unawareness setting (Eq. (P_∞)) reduces, in the awareness setting, to the problem given in Eq. (1). For this reason, we focus in the remainder of the paper on the more general unawareness setting, noting that the awareness case follows as a direct special case of our results.

Remark 2.1 (Noisy attributes). Our formulation also covers the case where a proxy $\hat{S}$ is available at inference time. Indeed, including $\hat{S}$ as a covariate in the feature vector X, we obtain $\Delta(x, \hat{s}) = \frac{\mathbb{P}(S=+|X=x, \hat{S}=\hat{s})}{p^+} - \frac{\mathbb{P}(S=-|X=x, \hat{S}=\hat{s})}{p^-}$, thus interpolating between the two settings. When $\hat{S}$ is X-measurable or contains uninformative noise, $\mathbb{P}(S|X, \hat{S}) = \mathbb{P}(S|X)$, recovering the unaware formulation. Conversely, when $\hat{S} = S$, we recover the aware setting [11].

2.2 Optimal Transport for relaxed fairness: a unified framework

In many practical settings, enforcing DP can significantly degrade predictive accuracy, sometimes to an unacceptable extent. In this section, we present our general framework for relaxed fair regression. As established in Section 2.1, DP is satisfied if and only if the pushforwards of the positive and negative variations of the signed measure $\chi^+ - \chi^-$ are equal. To achieve relaxed fairness, we penalize the risk of a prediction function by its $\mathcal{D}$-unfairness defined below.

Definition 1 ($\mathcal{D}$-unfairness)**.** Let $\mathcal{D}$ be a probability discrepancy measure over $\mathcal{P}(\mathbb{R})$, and let $m = \mathrm{TV}(\chi^+, \chi^-)/2$. The $\mathcal{D}$-unfairness of a prediction function f is defined as

$$\mathcal{U}_\mathcal{D}(f) = m \cdot \mathcal{D}\left(f_\#(\chi^+ - \chi^-)_+,\ f_\#(\chi^+ - \chi^-)_-\right).$$

We consider two choices for the discrepancy measure $\mathcal{D}$: the squared Wasserstein-2 distance ($\mathcal{U}_{\mathcal{W}_2^2}$), which evaluates the quadratic displacement required to align the distributions, and the Total Variation distance ($\mathcal{U}_{TV}$), which evaluates the maximal probability mass of individuals receiving disparate predictions. Both penalties can be expressed as $\mathcal{D}(\nu^+, \nu^-) = \mathrm{OT}_{c_\mathcal{D}}(\nu^+, \nu^-)$ for some cost function $c_\mathcal{D}$, equal to $c_2(y, y') = (y - y')^2$ for $\mathcal{W}_2^2$ and to $c_0(y, y') = \mathbb{1}_{\{y \neq y'\}}$ for TV. Our objective is to find the almost-fair prediction function $f^*_{\lambda,\mathcal{D}}$ solving the following penalized risk minimization problem:

$$f^*_{\lambda,\mathcal{D}} \in \arg\min_f\ \mathbb{E}\left[\left(f(X) - \eta(X)\right)^2\right] + \lambda\, \mathcal{U}_\mathcal{D}(f). \tag{3}$$

On the one hand, driving $\lambda \to \infty$ enforces $\mathcal{U}_\mathcal{D}(f) = 0$, which is equivalent to enforcing DP. On the other hand, the choice of unfairness measure $\mathcal{U}_\mathcal{D}$ enables us to cast the relaxed regression problem as an OT problem for general values of λ. In particular, it upper bounds the $\mathcal{D}$ distance between the distributions of the predictions: $\mathcal{U}_\mathcal{D}(f) \geq \mathcal{D}\left(\mathcal{L}(f(X)|S = +), \mathcal{L}(f(X)|S = -)\right)$, with equality if $\mathcal{D} = \mathrm{TV}$, as well as in the awareness setting (see Appendix A.1). When $\mathcal{D} = \mathcal{W}_2$, the gap is still bounded by the TV distance between the pushforward variations (see Appendix A.1). Moreover, our algorithm empirically promotes low unfairness with respect to multiple metrics (see Section 4).

Relaxed fair regression as an OT problem. To solve (3), we adopt an OT perspective. Under Assumption 1, we show that when $\mathcal{D}$ is taken to be either $\mathcal{W}_2^2$ or TV, and the cost $c_\mathcal{D}$ is such that $\mathcal{D}(\nu^+, \nu^-) = \mathrm{OT}_{c_\mathcal{D}}(\nu^+, \nu^-)$, the value of (3) coincides with that of the following OT problem

$$\inf_{\nu^+, \nu^- \in \mathcal{P}(\mathbb{R})} \left\{ \mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \nu^+) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \nu^-) + \lambda\, \mathrm{OT}_{c_\mathcal{D}}(\nu^+, \nu^-) \right\}. \tag{$P_{\lambda,\mathcal{D}}$}$$

Theorem 1. *Let $\mathcal{D}(\cdot,\cdot) = \mathrm{OT}_{c_\mathcal{D}}(\cdot,\cdot)$ for $c_\mathcal{D}(y,y') = (y-y')^2$ or $c_\mathcal{D}(y,y') = \mathbb{1}_{\{y\neq y'\}}$. Under Assumption 1, Problem $(P_{\lambda,\mathcal{D}})$ is solved by deterministic transport maps $T^+_{\lambda,\mathcal{D}}$ and $T^-_{\lambda,\mathcal{D}}$ which push $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ to the minimizing measures ν^+ and ν^-, respectively. Moreover, the solution of the relaxed fair regression problem* (3) *is given by*

$$f^*_{\lambda,\mathcal{D}}(x) = \begin{cases} T^+_{\lambda,\mathcal{D}}(\eta(x),\Delta(x)) & \text{if} \quad \Delta(x) > 0, \\ T^-_{\lambda,\mathcal{D}}(\eta(x),\Delta(x)) & \text{if} \quad \Delta(x) < 0, \\ \eta(x) \quad \text{else}. \end{cases} \tag{4}$$

Theorem 1 shows that the relaxed fair regression problem reduces to solving Problem $(P_{\lambda,\mathcal{D}})$ and finding an OT map. The corresponding problems arising in the unawareness and awareness settings are illustrated in Fig. 2. In the awareness setting, the solution amounts to constructing transport maps between one-dimensional distributions. In the unawareness setting, by contrast, one must transport two-dimensional measures $\mu^\pm_{\eta,\Delta}$ to a one dimensional prediction, where each point (h,d) in the support of $\mu^\pm_{\eta,\Delta}$ encodes the Bayes prediction η and the signed group probability Δ. Crucially, we show in Props. 1 and 2 that the 2-to-1-dimensional barycenter problem can be reduced to an OT problem between two 2-dimensional measures, with a penalty-dependent ground cost (C_λ), thereby enabling efficient numerical estimation.

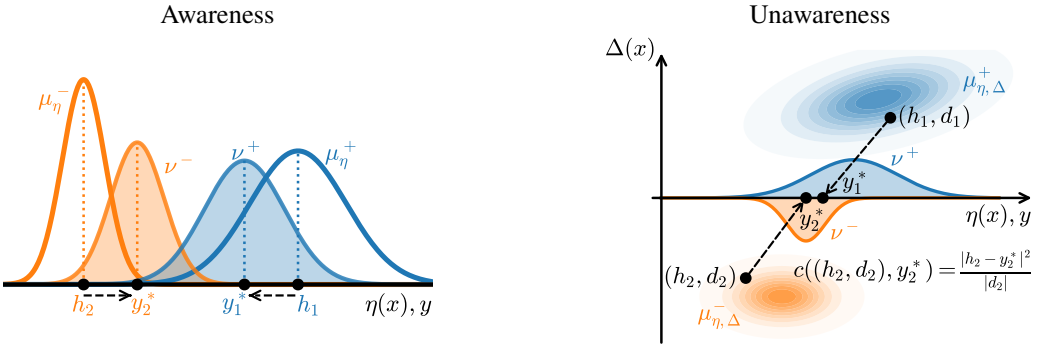

Figure 2: **Geometry of the OT maps.** *Left (awareness):* 1D-to-1D transport mapping group-conditional distributions to almost-fair predictions. *Right (unawaress):* 2D-to-1D transport mapping the joint distributions of $(\eta(X), \Delta(X))$ under $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ to almost-fair predictions.

In Table 1, we instantiate the general framework in $P_{\lambda,\mathcal{D}}$ for the aware and unaware settings, and for both unfairness measures considered. We obtain three new formulations for relaxed fair regression, and we recover one existing formulation **FAUST-A** $\mathcal{W}_2$ [11].

2.3 New results for relaxed fairness

Theorem 1 establishes that the solution to the fair relaxation problem is given by the OT maps solving $(P_{\lambda,\mathcal{D}})$. In this section, we show that $(P_{\lambda,\mathcal{D}})$ can itself be reduced to an OT problem between $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ with an appropriately chosen cost. Specializing to the $\mathcal{W}_2^2$ and the TV-fairness penalty, we obtain solutions to (3) expressed explicitly in terms of the OT maps. Closed-form solutions in the awareness setting follow as a corollary.

Relaxation under $\mathcal{W}_2^2$-fairness. We consider the fairness penalty $\mathcal{D}(\cdot,\cdot) = \mathcal{W}_2^2(\cdot,\cdot)$. The following proposition reduces $(P_{\lambda,\mathcal{W}_2})$ to an OT problem between $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ for a suitable cost, and gives an expression for the optimal almost-fair prediction via the OT plan. The proof is in Section A.4.

Proposition 1 (**FAUST-U** $\mathcal{W}_2$)**.** *Under Assumption 1, Problem $(P_{\lambda,\mathcal{W}_2})$ is equivalent to the OT problem*

$$\pi^* \in \arg\min_{\pi\in\Pi(\mu^+_{\eta,\Delta},\mu^-_{\eta,\Delta})} \int_{\mathbb{R}^2\times\mathbb{R}^2} C^u_{\lambda,c_2}(z_1,z_2)\,\mathrm{d}\pi(z_1,z_2), \tag{5}$$

where $C^u_{\lambda,c_2}(z_1,z_2) = \frac{\lambda}{1+\lambda(|d_1|+|d_2|)}(h_1-h_2)^2$ for $z_1 = (h_1,d_1)$ and $z_2 = (h_2,d_2)$. Moreover, a solution to the almost-fair regression problem in (3) *is given by Eq.* (4) *with*

	Unaware Setting	Aware Setting
Objective	$\mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta},\nu^+) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta},\nu^-)$ $+\lambda\,\mathrm{OT}_{c_{\mathcal{D}}}(\nu^+,\nu^-)$	$p^+\,\mathrm{OT}_{c_2}(\mu^+_{\eta},\nu^+) + p^-\,\mathrm{OT}_{c_2}(\mu^-_{\eta},\nu^-)$ $+\lambda\,\mathrm{OT}_{c_{\mathcal{D}}}(\nu^+,\nu^-)$
$\mathcal{W}_2^2$**-fairness**	**FAUST-U** $\mathcal{W}_2$ (Prop. 1) **Cost:** $c_u(z,y) = (h-y)^2/\lvert d\rvert$ **Penalty:** $c_{\mathcal{D}}(y,y') = (y-y')^2$ **Ground cost:** $C^u_{\lambda,c_2} = \frac{\lambda(h_1-h_2)^2}{1+\lambda(\lvert d_1\rvert+\lvert d_2\rvert)}$	**FAUST-A** $\mathcal{W}_2$ (Chzhen and Schreuder, Cor. 1) **Cost:** $c_2(y,y') = (y-y')^2$ **Penalty:** $c_{\mathcal{D}}(y,y') = (y-y')^2$ **Ground cost:** $C^a_{\lambda,c_2} = \frac{\lambda p^+p^-}{\lambda+p^+p^-}(h_1-h_2)^2$
TV-fairness	**FAUST-U TV** (Prop. 2) **Cost:** $c_u(z,y) = (h-y)^2/\lvert d\rvert$ **Penalty:** $c_{\mathcal{D}}(y,y') = \mathbb{1}_{\{y\neq y'\}}$ **Ground cost:** $C^u_{\lambda,c_0} = \min(\lambda, \frac{(h_1-h_2)^2}{\lvert d_1\rvert+\lvert d_2\rvert})$	**FAUST-A TV** (Cor. 2) **Cost:** $c_2(y,y') = (y-y')^2$ **Penalty:** $c_{\mathcal{D}}(y,y') = \mathbb{1}_{\{y\neq y'\}}$ **Ground cost:** $C^a_{\lambda,c_0}=\min(\lambda, p^+p^-(h_1-h_2)^2)$

Table 1: Summary of the unified OT framework. The general problem $P_{\lambda,\mathcal{D}}$ instantiates into specific pointwise closed-form reductions depending on error cost $c \in \{c_u, c_2\}$ and penalty $c_{\mathcal{D}}$.

$$T^+_{\lambda,\mathcal{W}_2^2}(z_1) = \frac{(1+\lambda|d_2|)h_1 + \lambda|d_1|h_2}{1+\lambda(|d_1|+|d_2|)} \quad \textit{and} \quad T^-_{\lambda,\mathcal{W}_2^2}(z_2) = \frac{\lambda|d_2|h_1 + (1+\lambda|d_1|)h_2}{1+\lambda(|d_1|+|d_2|)} \tag{6}$$

for all pairs (z_1, z_2) *in the support of* π^*.

Prop. 1 provides a roadmap for building almost-fair classifiers: first, solve the OT problem (5) with ground cost C_{λ,c_2}, then compute the prediction function via the OT plan using Eq. (4) and Eq. (6). While Eq. (5) is solved by a (non-deterministic) transport plan, we show in Section A.4 that the value defining $T^+_{\lambda,\mathcal{W}_2^2}(z_1)$ does not depend on the choice of z_2 for (z_1, z_2) in the support of π^*, so that $T^+_{\lambda,\mathcal{W}_2^2}$ is a well-defined map (and similarly for $T^-_{\lambda,\mathcal{W}_2^2}$). As noted in Section B and Remark A.1, this formulation encompasses prior results as boundary cases. In the exact fairness limit ($\lambda\to\infty$), Prop. 1 recovers, in the awareness framework, the fair regression function of Chzhen et al. [12], and in the unawareness framework, the OT map of Divol and Gaucher [16]. For generic $\lambda > 0$, in the awareness framework it recovers the geodesic interpolations of Chzhen and Schreuder [11], while in the unawareness framework it yields a new characterization of the optimal predictor under a relaxed fairness constraint.

Relaxation under TV-fairness We now consider the fairness penalty $\mathcal{D}(\cdot,\cdot) = \mathrm{TV}(\cdot,\cdot)$. The following proposition reduces $(P_{\lambda,\,\mathrm{TV}})$ to an OT problem between $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ with a suitable cost, and expresses the optimal almost-fair prediction in terms of the OT plan. The proof is in Section A.5.

Proposition 2 (FAUST-U TV). *Under Assumption 1, Problem* $(P_{\lambda,\,\mathrm{TV}})$ *is equivalent to the OT problem*

$$\pi^* \in \arg\min_{\pi\in\Pi(\mu^+_{\eta,\Delta},\mu^-_{\eta,\Delta})} \int_{\mathbb{R}^2\times\mathbb{R}^2} C^u_{\lambda,c_0}(z_1,z_2)\,\mathrm{d}\pi(z_1,z_2), \tag{7}$$

where $C^u_{\lambda,c_0}(z_1,z_2) = \min\left(\lambda,\ \frac{(h_1-h_2)^2}{|d_1|+|d_2|}\right)$. *Moreover, a solution to the fair regression problem in Eq.* (3) *is given by Eq.* (4) *with*

$$\begin{cases} T^+_{\lambda,\,\mathrm{TV}}(z_1) = T^-_{\lambda,\,\mathrm{TV}}(z_2) = \dfrac{|d_2|h_1 + |d_1|h_2}{|d_1|+|d_2|} & \textit{if} \quad \frac{(h_1-h_2)^2}{|d_1|+|d_2|} \le \lambda, \\ T^+_{\lambda,\,\mathrm{TV}}(z_1) = h_1, \quad T^-_{\lambda,\,\mathrm{TV}}(z_2) = h_2 & \textit{if} \quad \frac{(h_1-h_2)^2}{|d_1|+|d_2|} > \lambda, \end{cases} \tag{8}$$

for all pairs $z_1 = (h_1, d_1)$ *and* $z_2 = (h_2, d_2)$ *in the support of* π^*.

Here again, while Problem (7) is solved by a (non-deterministic) transport plan, we show in Section A.5 that $T^+_{\lambda,\,\mathrm{TV}}(z_1)$ is well-defined in the sense that its value does not depend on the choice of z_2 for (z_1, z_2) in the support of π^* (and similarly for $T^-_{\lambda,\,\mathrm{TV}}(z_2)$). In the exact fairness limit ($\lambda\to\infty$), Prop. 1 also recovers previous results [12, 16]. For $\lambda > 0$, it yields a novel characterization of the optimal predictor under a TV-fairness penalty in the unawareness framework.

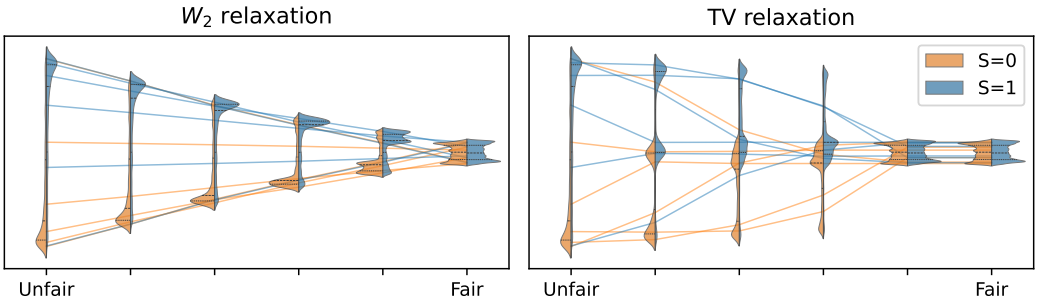

Figure 3: **Individual prediction trajectories for $\mathcal{W}_2$ vs. TV Relaxation.** Evolution of marginal distributions and subset of individual prediction paths from Unfair (Bayes) to exactly Fair (100%-0% of initial unfairness). The lines show $\mathcal{W}_2$ induces proportional shift for all individuals, while TV leaves predictions unchanged (horizontal lines) until a hard threshold triggers abrupt merging.

2.4 The implicit philosophies of relaxed fairness

The characterizations of the optimal almost-fair prediction function in Props. 1 and 2 highlight how the choice of fairness penalty impacts the behaviour of fair algorithms. In this section, we contrast the two philosophies induced by the geometry of the $\mathcal{W}_2$ and TV penalties.

Homogeneous gap reduction under Wasserstein relaxation. The $\mathcal{W}_2$ penalty penalizes the magnitude of the discrepancy between corresponding quantiles of the prediction in the two groups. This implies that it is preferable for all individuals to be slightly moved together toward the fair prediction than for a few to experience massive disparity. Philosophically, this aligns with an *egalitarian* view of fairness [35], which treats inequality as a shared societal burden to be mitigated globally across the population.

Heterogeneous parity under Total Variation relaxation. The TV penalty penalizes non-overlapping probability mass, ignoring distances in the prediction space. Compromising by moving predictions slightly closer yields no benefit unless they are merged perfectly. As shown in Prop. 2, this results in a hard-thresholding behavior. This implies that TV enforces parity where it is "affordable" (i.e., where the cost is low), and accepts total unfairness for the extremes of the distribution. TV ensures that a maximized subset of the population achieves indistinguishability. Ethically, this reflects a *sufficientarian* approach [22], focusing on securing a plateau of equality for the majority.

Illustration of the paradigms. We illustrate the effect of the $\mathcal{W}_2$ and TV relaxations on a synthetic dataset for the unaware setting (Fig. 3) using the implementation detailed in Section 3. The individual trajectories across relaxation levels reveal the distinct mechanisms of each penalty. As the fairness constraint tightens, the $\mathcal{W}_2$ penalty shifts the distributions uniformly, with individual predictions linearly interpolating toward the barycenter. By contrast, the TV penalty exhibits hard-thresholding behavior. Predictions either remain at their original values (horizontal paths) or merge. Furthermore, a saturation effect occurs: for sufficiently large λ, the solution stabilizes to the fair prediction. For data generation details, visualizations of the mapping geometries, and further examples, see Section F.1.

3 Algorithmic implementation

In the previous section, we expressed the solution of the almost-fair regression problem in terms of an OT problem between population distributions. Building on this characterization, we now propose an algorithm for **F**airness for **A**ware and **U**naware **Se**T**t**ings (**FAUST**). We assume access to an estimator $\hat{\eta}$ of the Bayes regression function η and an estimator $\hat{\mathbb{P}}(S = s|X = x)$ of the posterior group probabilities $\mathbb{P}(S = s|X = x)$ (which can be obtained using standard regression and probabilistic classification methods), and to an unlabeled dataset $(x_i, s_i)_{i\leq n}$. The pseudocode is in Section D.

Estimation of $\hat{\Delta}(x)$. We first estimate $\hat{p}^+$, $\hat{p}^-$ of p^+ and p^- using $(x_i, s_i)_{i\leq n}$, and define $\hat{\Delta}(x) = \frac{\hat{\mathbb{P}}(S=+|X=x)}{\hat{p}^+} - \frac{\hat{\mathbb{P}}(S=-|X=x)}{\hat{p}^-}$. We then apply a threshold τ to partition training indices into positive ($\mathcal{I}^+ = \{i : \hat{\Delta}(x_i) > \tau\}$) and negative ($\mathcal{I}^- = \{i : \hat{\Delta}(x_i) < -\tau\}$) subsets based on the sign

of $\hat{\Delta}$, forming the discrete supports of our empirical measures. Crucially, to account for finite-sample estimation errors, we normalize the local weights ($\hat{d}_i \propto \hat{\Delta}(x_i)$ for $i \in \mathcal{I}^+$ and $\hat{d}_i \propto -\hat{\Delta}(x_i)$ for $i \in \mathcal{I}^-$) to enforce mass conservation ($\sum_{i\in\mathcal{I}^+} |\hat{d}_i| = \sum_{i\in\mathcal{I}^-} |\hat{d}_i| = 1$).

OT plan and barycentric projection. From the measures $(\hat{\eta}_i, \hat{d}_i)_{i\in\mathcal{I}^+}$ and $(\hat{\eta}_j, \hat{d}_j)_{j\in\mathcal{I}^-}$, we construct the pairwise cost matrix $\mathbf{C}_\lambda \in \mathbb{R}_+^{|\mathcal{I}^+|\times|\mathcal{I}^-|}$ using the closed-form solutions derived in Prop. 1 (for the $\mathcal{W}_2$ relaxation) or Prop. 2 (for the TV relaxation). We then solve the discrete OT problem between the weighted measures $(\hat{\eta}_i, \hat{d}_i)_{i\in\mathcal{I}^+}$ and $(\hat{\eta}_j, \hat{d}_j)_{j\in\mathcal{I}^-}$ with corresponding weights $(\hat{d}_i)_{i\in\mathcal{I}^+}$ and $(\hat{d}_i)_{i\in\mathcal{I}^+}$ to obtain the optimal coupling matrix $\mathbf{\Pi}$. This reweighting allows us to estimate the distribution $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$, as discussed in Section D. With a slight abuse of notation, let $T^s_{\lambda,\mathcal{D}}(z_i, z_j)$ denote the closed-form target mapping evaluated for the coupled pair (z_i, z_j) (derived via Eq. (6) for $\mathcal{W}_2$ and Eq. (8) for TV). The pseudo-label for point i is then computed as

$$\hat{f}_i = \frac{1}{|\hat{d}_i|} \sum_{j\in\mathcal{I}^{(-s)}} \mathbf{\Pi}_{i,j} T^s_{\lambda,\mathcal{D}} \left((\hat{\eta}_i, \hat{d}_i), (\hat{\eta}_j, \hat{d}_j) \right).$$

Continuous mapping and final predictor. For out-of-sample predictions, we concatenate training points $(x_i)_{i\le n}$ and fit a regressor $\hat{f}_{fair}$ to predict $(\hat{f}_i)_{i\le n}$. The prediction for a new instance x is obtained by evaluating $\hat{\eta}(x)$ and $\hat{\Delta}(x)$, and predicting $\hat{f}_{\text{fair}}(\hat{\eta}(x), \hat{\Delta}(x))$.

Computational complexity. The complexity is dominated by the $O(n^3 \log n)$ discrete OT step, computed using the network simplex implementation from the POT library [21]. Despite this worst-case bound, our approach is more efficient than Taturyan et al. [38] as it bypassed stochastic gradient descent. Empirically, it runs in seconds for $n \sim 10^4$, significantly outperforming FairReg, and can be further accelerated via mini-batching [18] to scale to millions of samples without degrading accuracy or fairness (see our extensive benchmark in Section H)

4 Numerical experiments

4.1 Baseline and state-of-the-art comparison

We benchmark our methods against the unconstrained (and unfair) Empirical Risk Minimizer (ERM) and the state-of-the-art methods for unaware relaxed fair regression, FairReg [38] and FRAPPÉ [14].

Unaware Baselines. FairReg relies on a dual stochastic convex program to learn a randomized prediction rule over a discretized finite grid. FRAPPÉ relies on stochastic gradient descent to learn an additive post-hoc neural network correction that minimizes a fairness-regularized loss. Full algorithmic details and implementation parameters for both baselines are deferred to Section E.1. We compare our FAUST method using the experimental setup of FairReg.

Aware and proxy-based baselines. To establish the empirical lower bound for the accuracy-fairness trade-off, we include the aware OT models (FAUST-A $\mathcal{W}_2$ and FAUST-A TV), which have access to the true sensitive attribute S at both training and inference time. We also benchmark against a naive plug-in estimator (Plug-In $\mathcal{W}_2$). A common heuristic [4, 9, 28] for the unaware setting is to predict a proxy $\hat{S}$ using a probabilistic classifier and apply the classical 1D aware OT mapping. Comparing these approaches highlights the cost of classification errors compared to our 2D framework. A common heuristic [4, 9, 28] for the unaware setting is to predict a proxy $\hat{S}$ using a probabilistic classifier and apply the classical 1D aware OT mapping. Comparing these approaches highlights the cost of classification errors compared to our 2D framework.

Evaluation metrics. We evaluate predictive utility via Mean Squared Error (MSE). To quantify fairness, we report three metrics. First, the Wasserstein-2 ($\mathcal{W}_2$) and Total Variation (TV) unfairness. Then, to ensure fair comparison with the FairReg baseline, we report $\text{KS}^{\mathcal{T}}$, which approximates the KS distance by computing the maximum absolute c.d.f. difference across a discrete 50-bin grid $\mathcal{T}$ (we demonstrate the robustness of our performance to this choice of discretization in Section G.5).

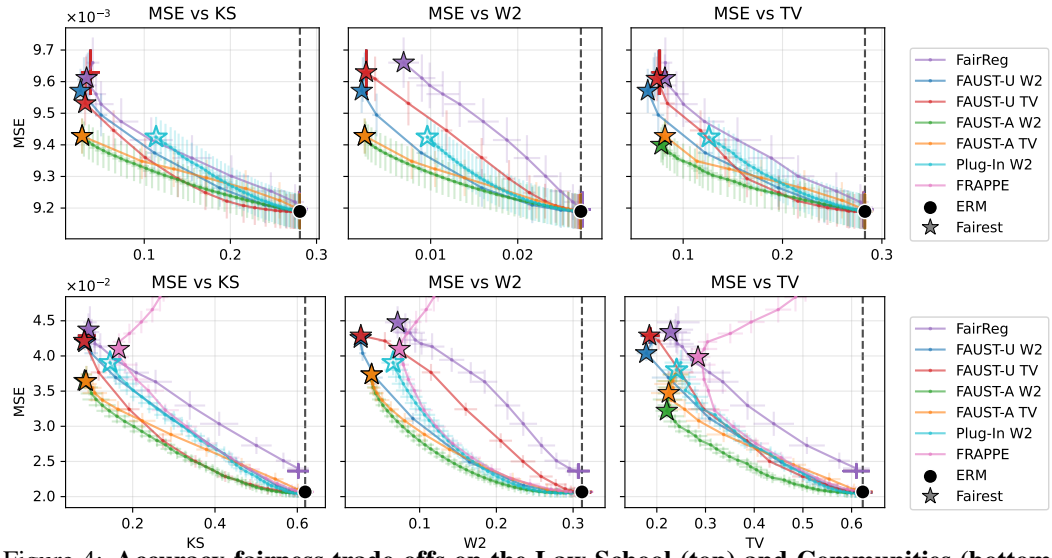

Figure 4: **Accuracy-fairness trade-offs on the Law School (top) and Communities (bottom) datasets.** Relaxation trajectories are shown from the unconstrained ERM (black dot) to the exact fair models (stars). The MSE of the constant predictor is 10.1×10^{-3} on Law School, and 5.5×10^{-2} on Communities. Error bars indicate the standard deviation across 10 random data splits. FRAPPÉ is cropped when it falls outside the axis limits; full trajectories are provided in Section G.2.

4.2 Real-world datasets

To evaluate our framework against the FairReg baseline [38], we use real-world datasets from their evaluation protocol: Communities and Crime, and Law School Admissions. Identical experiments and visualizations for a synthetic and the Adult (Census Income) dataset are provided in Sections F.2 and G respectively, which show similar performances of our OT algorithms.

Datasets descriptions. The Law School Admissions dataset comprises 20,649 student records from 163 U.S. law schools, featuring academic indicators (e.g., LSAT, GPA) and demographic variables. The regression task is to predict the student's standardized first-year GPA (`zfygpa`). Because the data exhibits racial imbalance (approximately 84.6% White), we follow standard practice and define the binary sensitive attribute S as White ($S = 1$) versus Non-White ($S = 0$).
The Communities and Crime dataset comprises 1,994 U.S. communities characterized by 128 socio-economic and demographic features. The regression task is to predict the violent crime rate per 100,000 population (`ViolentCrimesPerPop`). To construct S, we identify the majority racial demographic within each community and binarize it into White ($S = 1$) versus non-White ($S = 0$).

Results. As shown in Section 4.2, our unaware OT framework (FAUST-U) consistently matches or outperforms the baselines. The naive estimator (Plug-In W2) degrades the accuracy-fairness trade-off by relying on a noisy proxy attribute, pushing its curve strictly worse than FairReg on the Communities dataset. Furthermore, the gradient-based FRAPPÉ baseline suffers from severe optimization instability: as the fairness constraint tightens, MSE degrades while failing to reach parity (see Section G.2 for full plots).
Conversely, by relying on exact geometric mappings rather than stochastic approximations, FAUST-U $\mathcal{W}_2$ tracks the theoretical aware bounds. It achieves superior trade-offs on $\mathcal{W}_2$ and remains competitive on KS and TV. Interestingly, $\mathcal{W}_2$-penalized models achieve better tradeoffs than TV-based ones *even when evaluated on the TV metric*. Finally, our framework exhibits empirical robustness to finite-sample errors and miscalibration of the base estimators, as detailed in Section G.4.

Extended empirical validation. To validate our framework, we conduct an extensive set of supplementary experiments detailed in the Appendix. We provide complete evaluations on the Adult dataset (Section G) with trade-off plots, detail the implementations of the baselines (Sections E.1 and G.2), and illustrate the mapping geometries across varying levels of confounding (Sections F.1 and F.2). We also stress-test the framework's stability, confirming its robustness to finite-sample estimation errors, artificial noise injected into the base estimators (Section G.4), and varying metric discretization grids (Section G.5). Furthermore, we demonstrate that FAUST-U is model-agnostic,

consistently maintaining its superior accuracy-fairness trade-offs across different unconstrained base estimators, including Random Forests and Gradient Boosting (Section G.6). Finally, we expand on our computational efficiency results: FAUST-U runs up to $28\times$ faster than FairReg and scales to massive datasets (e.g., $N = 200,000$) in seconds via Minibatch OT, without degrading parity (Section H).

5 Limitations and perspectives

This work introduces a unified geometric framework for relaxed fair regression, yielding closed-form optimal predictors via OT maps. This characterization clarifies the philosophical implications of fairness penalties and enables a highly efficient, state-of-the-art post-processing algorithm for both aware and unaware settings.

We identify three main limitations. First, our closed-form reductions are restricted to TV and $\mathcal{W}_2^2$ penalties. While these provide a geometric lens to contrast sufficientarian and egalitarian philosophies, extending the framework to general OT_c costs could inspire novel fair algorithms. Second, extending our binary sensitive attribute formulation to multi-group fairness requires generalizing the Jordan decomposition to multi-marginal OT, posing an interesting open challenge. Finally, as a post-processing method, performance depends on the base estimators η and Δ. While proving non-asymptotic convergence rates for plug-in OT maps is notoriously challenging [34], our framework bypasses the curse of dimensionality. By mapping the high-dimensional feature space $\mathcal{X}$ into a 2D pushforward space $(\eta(X), \Delta(X))$, the empirical 2D-to-1D OT problem converges at an almost parametric rate [2]. Thus, given consistent base estimators, our algorithm should approximate the optimal solution at a dimension-independent statistical rate, with formal theoretical guarantees left for future work.

Acknowledgements and Disclosure of Funding

This work was supported by Hi! PARIS and the ANR/France 2030 program (ANR-23-IACL-0005), as well as by the Fondation de l'Ecole Polytechnique. Additional support was provided by the French National Research Agency (ANR) through grant ANR-23-CE23-0002, the PEPR IA FOUNDRY project (ANR-23-PEIA-0003), the France 2030 program under the MacLeOD project (ANR-25-PEIA-0005) and the Labex Ecodec (ANR-11-LABEX-0047).

References

[1] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

[2] Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4(4): 259–264, 1984.

[3] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin Heidelberg, 3 edition, 2006. ISBN 978-3-540-32696-0. doi: 10.1007/3-540-29587-9.

[4] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1770–1780. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/awasthi20a.html.

[5] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1770–1780. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/awasthi20a.html.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[7] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. In *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, ICML Workshop, August 2017. URL https://arxiv.org/abs/1706.02409. Co-located with KDD 2017. Preprint: arXiv:1706.02409.

[8] Beepul Bharti, Paul Yi, and Jeremias Sulam. Estimating and controlling for equalized odds via sensitive attribute predictors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=e2aCgjtjMR.

[9] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 339–348. ACM, January 2019. doi: 10.1145/3287560.3287594. URL http://dx.doi.org/10.1145/3287560.3287594.

[10] Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. A general approach to fairness with optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3633–3640, Apr. 2020. doi: 10.1609/aaai.v34i04.5771. URL https://ojs.aaai.org/index.php/AAAI/article/view/5771.

[11] Evgenii Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, August 2022. doi: 10.1214/22-AOS2198. URL https://arxiv.org/abs/2007.14265.

[12] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

[13] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. URL http://jmlr.org/papers/v20/18-616.html.

[14] Alexandru Ţifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FrappÉ: a group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[15] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25 (130):1–46, 2024. URL https://jmlr.org/papers/v25/23-0322.html.

[16] Vincent Divol and Solenne Gaucher. Demographic parity in regression and classification within the unawareness framework, 2024. URL https://arxiv.org/abs/2409.02471.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226. ACM, January 2012. doi: 10.1145/2090236.2090255. URL https://arxiv.org/abs/1104.3913.

[18] Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties. In *AISTATS 2020 - 23nd International Conference on Artificial Intelligence and Statistics*, volume volume 108 of *PMLR*, pages 1–20, Palermo, Italy, June 2020. URL https://hal.science/hal-02502329.

[19] Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/fatras20a.html.

[20] Eugene A Feinberg, Pavlo O Kasyanov, and Nina V Zadoianchuk. Fatou's lemma for weakly converging probabilities. *Theory of Probability & Its Applications*, 58(4):683–689, 2014.

[21] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[22] Harry Gordon Frankfurt. Equality as a moral ideal. *Ethics*, 98(1):21–43, October 1987. doi: 10.1086/292913. URL https://www.jstor.org/stable/2381290.

[23] Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with wasserstein barycenters for non-decomposable performance measures. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2436–2459. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/gaucher23a.html.

[24] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning, 2020. URL https://arxiv.org/abs/2005.11720.

[25] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness, 2018. URL https://arxiv.org/abs/1806.11212.

[26] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hashimoto18a.html.

[27] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Manage. Sci.*, 68(3):1959–1981, March 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2020.3850. URL https://doi.org/10.1287/mnsc.2020.3850.

[28] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Manage. Sci.*, 68(3):1959–1981, March 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2020.3850. URL https://doi.org/10.1287/mnsc.2020.3850.

[29] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/07fc15c9d169ee48573edd749d25945d-Paper.pdf.

[30] Thibaud Leteno, Michael Perrot, Charlotte Laclau, Antoine Gourru, and Christophe Gravier. Fair text classification via transferable representations. *Journal of Machine Learning Research*, 26(239):1–47, 2025. URL http://jmlr.org/papers/v26/25-0485.html.

[31] D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc., 2009. ISBN 9780821886274. URL https://books.google.fr/books?id=6Cg5Nq5sSv4C.

[32] Zhu Li, Adrian Perez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness, 2019. URL https://arxiv.org/abs/1911.04322.

[33] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf.

[34] Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps, 2024. URL https://arxiv.org/abs/2107.12364.

[35] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[36] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser Cham, 1 edition, 2015. ISBN 978-3-319-20827-5. doi: 10.1007/978-3-319-20828-2. Published 27 October 2015.

[37] Eloi Tanguy, Julie Delon, and Nathaël Gozlan. Computing barycentres of measures for generic transport costs. *arXiv preprint arXiv:2501.04016*, 2024.

[38] Gayane Taturyan, Evgenii Chzhen, and Mohamed Hebiri. Regression under demographic parity constraints via unlabeled post-processing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 117917–117953. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d5c3ecf397fff63419bb5f5f2d8afe33-Paper-Conference.pdf.

[39] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media, 2008. ISBN 9783540710509.

[40] Mariia Vladimirova, Federico Pavone, and Eustache Diemert. Fairjob: A real-world dataset for fairness in online systems. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=lygceqe21t.

[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL `http://jmlr.org/papers/v20/18-262.html`.

Appendix Table of Contents

A Proofs of propositions and lemmas **2**

A.1 Tightness of the upper bound between $\mathcal{U}_{\mathcal{D}}$ and the conditional disparity 2

A.2 Equivalence between (3) and $(P_{\lambda,\mathcal{D}})$, assuming deterministic optimal transport maps . 2

A.3 Reduction of $(P_{\lambda,\mathcal{D}})$ to an OT problem between $\mu^{+}_{\eta,\Delta}$ and $\mu^{-}_{\eta,\Delta}$ 4

A.4 Proof of Prop. 1 (FAUST-U $\mathcal{W}_2$) 5

A.5 Proof of Prop. 2 (FAUST-U TV) 8

B Proofs for the relaxed aware setting **10**

B.1 Aware $\mathcal{W}_2$ relaxation 10

B.2 Aware TV relaxation 11

C Unfairness of plug-in approaches **12**

D Algorithm pseudocode **13**

E Experimental setup and implementation details **14**

E.1 Details on the Unaware Baselines (FairReg and FRAPPÉ) 14

E.2 General experimental details 15

F Experiments on synthetic datasets **16**

F.1 Empirical illustration of the geometric relaxations 16

F.2 Impact of confounding strenght and trade-offs 17

G Experiments on real-world datasets **18**

G.1 Adult dataset 18

G.2 Accuracy-Fairness trade-off 19

G.3 Evaluation of the fully constrained models 20

G.4 Robustness to estimation error 21

G.5 Sensitivity of fairness metrics to discretization grids 25

G.6 Robustness across different base estimators 25

H Computational efficiency **28**

A Proofs of propositions and lemmas

Notation for the proofs. Throughout the appendix, we define the 2D mapped feature space as $\mathcal{Z} := \mathbb{R} \times \mathbb{R}$. An element in this space is denoted as $z = (h, d) \in \mathcal{Z}$, where $h \in \mathbb{R}$ represents the initial base prediction $\eta(x)$, and $d \in \mathbb{R}$ represents the local density ratio $\Delta(x)$. For readability, we denote its absolute value as $a = |d|$. Additionally, for a measure $\mu \in \mathcal{P}(\mathbb{R})$, we denote its generalized quantile function as $Q_\mu(t) = \inf\{y \in \mathbb{R} : F_\mu(y) \geq t\}$ for $t \in (0, 1)$.

We begin the proof of Theorem 1 by showing that if Problem $(P_{\lambda,\mathcal{D}})$ is solved by transport maps, then the solution to Problem (3) is indeed given as in (4). This is done in Lemma 1. The remainder of the proof will then consist in proving that for the choice of fairness penalties considered, and under Assumption 1, Problem $(P_{\lambda,\mathcal{D}})$ is indeed solved by transport maps. To do this, we show that Problem $(P_{\lambda,\mathcal{D}})$ reduces to an OT problem between measures $\mu^+_{\eta,\Delta}$, $\mu^-_{\eta,\Delta}$ (Lemma 2), before showing in Section A.4 and Section A.5 that for the $\mathcal{W}_2^2$ and the TV fairness penalties respectively, this problem is indeed solved by deterministic transport maps. Along the way, we prove Prop. 1 in Section A.4, and Prop. 2 in Section A.5.

A.1 Tightness of the upper bound between $\mathcal{U}_\mathcal{D}$ and the conditional disparity

In Definition 1, we defined the fairness penalty $\mathcal{U}_\mathcal{D}$ via the normalized variations of the demographic distributions rather than the true conditional disparity. We demonstrate here that the gap between these two quantities is either zero or tightly bounded.

Let $\nu^+ = \frac{1}{m} f_\sharp(\chi^+ - \chi^-)_+$ and $\nu^- = \frac{1}{m} f_\sharp(\chi^+ - \chi^-)_-$ be the normalized pushforwards. Let $a' = f_\sharp \chi^+$ and $b' = f_\sharp \chi^-$. Observe that $a' \geq m\nu^+$ and $b' \geq m\nu^-$, with the shared mass being $\sigma := a' - m\nu^+ = b' - m\nu^-$. We want to compare our proxy $\mathcal{U}_\mathcal{D}(f) = m\mathcal{D}(\nu^+, \nu^-)$ to the true disparity $\mathcal{D}(a', b')$.

First, if $\mathcal{D}$ is a norm induced by a distance (as is the case for the Total Variation distance), the gap is zero. Indeed, since σ cancels out, we have $m(\nu^+ - \nu^-) = a' - b'$. Because TV scales linearly with mass, $\mathcal{U}_{\mathrm{TV}}(f) = m\,\mathrm{TV}(\nu^+, \nu^-) = \mathrm{TV}(a', b')$.

A gap can only exist for the $\mathcal{W}_2$ penalty. In that case, we have $\mathcal{U}_{\mathcal{W}_2^2}(f) = m\mathcal{W}_2^2(\nu^+, \nu^-) \geq \mathcal{W}_2^2(a', b') = \mathcal{W}_2^2(\mathcal{L}(f(X)|S = +), \mathcal{L}(f(X)|S = -))$. This inequality naturally holds because a valid transport plan for the full distributions a' and b' can be constructed by leaving the shared mass σ stationary (incurring zero cost for this $1 - m$ fraction) and exclusively transporting the divergent mass $m\nu^+$ to $m\nu^-$. Since the true distance $\mathcal{W}_2^2(a', b')$ evaluates the infimum over all possible plans, its cost is upper-bounded by the cost of this specific construction, which is $m\mathcal{W}_2^2(\nu^+, \nu^-)$. A reverse inequality exists and takes the form

$$\mathcal{U}_{\mathcal{W}_2^2}(f) - \mathcal{W}_2^2(\mathcal{L}(f(X)|S = +), \mathcal{L}(f(X)|S = -)) \lesssim B^2 \min(m, 1 - m),$$

where m is the normalization constant defined earlier, and we assume that f takes its values in $[-B, B]$. This gap is zero in two extreme cases: if $m = 1$ (disjoint support case, where the sensitive attribute can be recovered perfectly from X) and if $m = 0$ (where predictions are fair). Moreover, this upper bound on the gap is computable in practice, since it relies only on m, a quantity evaluated during our metric computations.

Proof sketch. We use the dual formulation of OT: $\mathcal{W}_2^2(\mu, \nu) = \sup_{\phi \oplus \psi \leq |x-y|^2} \int \phi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu$. Let (ϕ, ψ) be an optimal pair for $(m\nu^+, m\nu^-)$. Since it is suboptimal for (a', b'), we have

$$m\mathcal{W}_2^2(\nu^+, \nu^-) - \mathcal{W}_2^2(a', b') \leq \int \phi \,\mathrm{d}(m\nu^+ - a') + \int \psi \,\mathrm{d}(m\nu^- - b') = \int g \,\mathrm{d}\sigma,$$

where $g = -\phi - \psi$. Optimal potentials ϕ and ψ inherit the modulus of continuity of the cost function. Here, the cost is $\mathcal{O}(B)$-Lipschitz, which implies that $|g| = \mathcal{O}(B^2)$. Thus, the gap is upper bounded by $\mathcal{O}(B^2)|\sigma| = \mathcal{O}(B^2)(1 - m)$. Moreover, $m\mathcal{W}_2^2(\nu^+, \nu^-) - \mathcal{W}_2^2(a', b') \leq m\mathcal{W}_2^2(\nu^+, \nu^-) \leq 4B^2 m$ since the cost is bounded by $4B^2$. □

A.2 Equivalence between (3) and $(P_{\lambda,\mathcal{D}})$, assuming deterministic optimal transport maps

Lemma 1. *If Problem $(P_{\lambda,\mathcal{D}})$ is solved by deterministic transport maps $T^+_{\lambda,\mathcal{D}}$ and $T^-_{\lambda,\mathcal{D}}$ which push $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ to the minimizing measures ν^+ and ν^-, respectively, then the solution of the relaxed*

fair regression problem (3) *is given by*

$$f^*_{\lambda,\mathcal{D}}(x) = \begin{cases} T^+_{\lambda,\mathcal{D}}(\eta(x), \Delta(x)) & \textit{if} \quad \Delta(x) > 0, \\ T^-_{\lambda,\mathcal{D}}(\eta(x), \Delta(x)) & \textit{if} \quad \Delta(x) < 0, \\ \eta(x) \quad \textit{else}. \end{cases}$$

Proof. Let us start by setting some additional notations. We denote $\chi = p^+\chi^+ + p^-\chi^-$ the law of X. Let $|\chi^+ - \chi^-|$ be the variation of $\chi^+ - \chi^-$, then for $m = \frac{\mathrm{TV}(\chi^+,\chi^-)}{2}$, we define the normalized Jordan decomposition of $\chi^+ - \chi^-$ through

$$\begin{cases} m(\chi^+ - \chi^-)_+ = \frac{1}{2}(|\chi^+ - \chi^-| + \chi^+ - \chi^-), \\ m(\chi^+ - \chi^-)_- = \frac{1}{2}(|\chi^+ - \chi^-| - \chi^+ + \chi^-), \end{cases}$$

that is $(m(\chi^+ - \chi^-)_+, m(\chi^+ - \chi^-)_-)$ is the Jordan decomposition of $\chi^+ - \chi^-$. Then, we have $\chi^+ - \chi^- = m(\chi^+ - \chi^-)_+ - m(\chi^+ - \chi^-)_-$, and

$$\begin{aligned} \Delta(x) &= \frac{\mathbb{P}(S = +|X)}{p^+} - \frac{\mathbb{P}(S = -|X)}{p^-} \\ &= \frac{\mathrm{d}(\chi^+ - \chi^-)}{\mathrm{d}\chi}(x) \\ &= m\frac{\mathrm{d}(\chi^+ - \chi^-)_+}{\mathrm{d}\chi}(x) - m\frac{\mathrm{d}(\chi^+ - \chi^-)_-}{\mathrm{d}\chi}(x). \end{aligned}$$

Finally, we also define $\mathcal{X}_+ = \{x \in \mathcal{X} : \Delta(x) > 0\}$, $\mathcal{X}_- = \{x \in \mathcal{X} : \Delta(x) < 0\}$, and $\mathcal{X}_= = \{x \in \mathcal{X} : \Delta(x) = 0\}$: $\mathcal{X}_+$ is the support of $(\chi^+ - \chi^-)_+$, and $\mathcal{X}_-$ is the support of $(\chi^+ - \chi^-)_-$. With these notations, for all prediction function f,

$$\begin{aligned} \mathbb{E}\left[(\eta(X) - f(X))^2\right] &= \int (\eta(x) - f(x))^2 \,\mathrm{d}\chi(x) \\ &= \int_{\mathcal{X}_+} (\eta(x) - f(x))^2 \frac{\mathrm{d}\chi}{\mathrm{d}(\chi^+ - \chi^-)_+}(x) \,\mathrm{d}(\chi^+ - \chi^-)_+(x) \\ &\qquad + \int_{\mathcal{X}_-} (\eta(x) - f(x))^2 \frac{\mathrm{d}\chi}{\mathrm{d}(\chi^+ - \chi^-)_-}(x) \,\mathrm{d}(\chi^+ - \chi^-)_-(x) \\ &\qquad + \int_{\mathcal{X}_=} (\eta(x) - f(x))^2 \,\mathrm{d}\chi(x) \\ &= \int_{\mathcal{X}_+} (\eta(x) - f(x))^2 \frac{m}{|\Delta(x)|} \,\mathrm{d}(\chi^+ - \chi^-)_+(x) \\ &\qquad + \int_{\mathcal{X}_-} (\eta(x) - f(x))^2 \frac{m}{|\Delta(x)|} \,\mathrm{d}(\chi^+ - \chi^-)_-(x) \\ &\qquad + \int_{\mathcal{X}_=} (\eta(x) - f(x))^2 \,\mathrm{d}\chi(x) \\ &= m \int_{\mathcal{X}_+} c_u((\eta(x), \Delta(x)), f(x)) \,\mathrm{d}(\chi^+ - \chi^-)_+(x) \\ &\qquad + m \int_{\mathcal{X}_-} c_u((\eta(x), \Delta(x)), f(x)) \,\mathrm{d}(\chi^+ - \chi^-)_-(x) \\ &\qquad + \int_{\mathcal{X}_=} (\eta(x) - f(x))^2 \,\mathrm{d}\chi(x). \end{aligned}$$

Recall that $\mu^+_{\eta,\Delta} = (\eta, \Delta)_\#(\chi^+ - \chi^-)_+$, and define $\tilde{\nu}^+ = f_\#(\chi^+ - \chi^-)_+$. Then, $(\eta(X), \Delta(X), f(X))$ defines a coupling between $\mu^+_{\eta,\Delta}$ and $\tilde{\nu}^+$, so

$$\int_{\mathcal{X}_+} c_u((\eta(x), \Delta(x)), f(x)) \,\mathrm{d}(\chi^+ - \chi^-)_+(x) \geq \mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \tilde{\nu}^+).$$

Defining similarly $\tilde{\nu}^- = f_\#(\chi^+ - \chi^-)_-$, we have

$$\int_{\mathcal{X}_-} c_u((\eta(x), \Delta(x)), f(x))\, \mathrm{d}(\chi^+ - \chi^-)_-(x) \geq \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \tilde{\nu}^-).$$

Now, by definition, $\mathcal{U}_\mathcal{D}(f) = m\, \mathrm{OT}_{c_\mathcal{D}}(\tilde{\nu}^+, \tilde{\nu}^-)$, so we find

$$\begin{aligned} &\min_f \mathbb{E}\left[(\eta(X) - f(X))^2\right] + \lambda \mathcal{U}_\mathcal{D}(f) \\ &\quad \geq m \inf_{\tilde{\nu}^+, \nu^- \in \mathcal{P}(\mathbb{R})} \Big(\mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \tilde{\nu}^+) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \tilde{\nu}^-) + \lambda\, \mathrm{OT}_{c_\mathcal{D}}(\tilde{\nu}^+, \tilde{\nu}^-) \Big). \end{aligned} \tag{9}$$

Now, if the OT problem on the right hand side of (9) is solved by deterministic transport maps $T^+_{\lambda,\mathcal{D}}$ and $T^-_{\lambda,\mathcal{D}}$ which push $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$ to the minimizing measures ν^+ and ν^-, respectively, then for $f^*_{\lambda,\mathcal{D}}$ given in Equation (4) we have

$$\begin{aligned} \int_{\mathcal{X}_+} c_u((\eta(x), \Delta(x)), f^*_{\lambda,\mathcal{D}}(x))\, \mathrm{d}(\chi^+ - \chi^-)_+(x) &= \mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \nu^+), \\ \int_{\mathcal{X}_+} c_u((\eta(x), \Delta(x)), f^*_{\lambda,\mathcal{D}}(x))\, \mathrm{d}(\chi^+ - \chi^-)_-(x) &= \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \nu^-), \\ \int_{\mathcal{X}_=} (\eta(x) - f^*_{\lambda,\mathcal{D}}(x))^2\, \mathrm{d}\chi(x) &= 0 \end{aligned}$$

so

$$\begin{aligned} &\mathbb{E}\left[(\eta(X) - f(X)^*)^2\right] + \lambda \mathcal{U}_\mathcal{D}(f^*) \\ &\qquad = m \min_{\tilde{\nu}^+, \tilde{\nu}^-} \Big(\mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \tilde{\nu}^+) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \tilde{\nu}^-) + \lambda\, \mathrm{OT}_{c_\mathcal{D}}(\tilde{\nu}^+, \tilde{\nu}^-) \Big). \end{aligned}$$

This concludes the proof of Lemma 1. □

A.3 Reduction of $(P_{\lambda,\mathcal{D}})$ to an OT problem between $\mu^+_{\eta,\Delta}$ and $\mu^-_{\eta,\Delta}$

As an introductory result, we use the Gluing lemma to show that the minimization problem $(P_{\lambda,\mathcal{D}})$ can be recast as a simple optimal transport problem for some appropriate cost.

Lemma 2. *Let $c_\mathcal{D} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be lower semi-continuous and bounded from below. For any pair of points $z_1 = (h_1, d_1)$ and $z_2 = (h_2, d_2)$, we define the pointwise relaxed cost*

$$C_{\lambda,\mathcal{D}}(z_1, z_2) := \inf_{y_1, y_2 \in \mathbb{R}} \left\{ \frac{(h_1 - y_1)^2}{|d_1|} + \frac{(h_2 - y_2)^2}{|d_2|} + \lambda c_\mathcal{D}(y_1, y_2) \right\}. \tag{10}$$

Then, the problem $(P_{\lambda,\mathcal{D}})$ is equivalent to the optimal transport problem using this pointwise cost, in the sense that

$$\inf(P_{\lambda,\mathcal{D}}) \;=\; \mathrm{OT}_{C_{\lambda,\mathcal{D}}}(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta}) \;=\; \min_{\pi \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})} \int_{\mathbb{R}\times\mathbb{R}} C_{\lambda,\mathcal{D}}(z_1, z_2)\, d\pi(z_1, z_2). \tag{11}$$

Moreover, let π^* be a minimizer of the right-hand side, and let $(y_1^*, y_2^*) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}^2$ be a measurable map such that $(y_1^*(z_1, z_2), y_2^*(z_1, z_2))$ realizes the infimum in (10) *for all $(z_1, z_2) \in \mathbb{R}^2 \times \mathbb{R}^2$. Then, $(\nu^{+*}, \nu^{-*}) = (y_1^*, y_2^*)_\sharp \pi_*$ is a minimizer of $(P_{\lambda,\mathcal{D}})$.*

Proof of Lemma 2. To prove this equivalence, we first establish that the pointwise optimization constitutes a lower bound for the global problem, and then demonstrate that this bound is exactly achieved.

Lower bound using the Gluing lemma. Let $\nu^+, \nu^- \in \mathcal{P}(\mathbb{R})$. The cost c_u is lower semi-continuous and bounded from below and we assume that the fairness penalty $c_{\mathcal{D}}$ is also lower semi-continuous and bounded from below. Therefore, by Theorem 1.5 from [36], optimal transport plans exist for all three costs c_u, $c_{\mathcal{D}}$ and $C_{\lambda,\mathcal{D}}$. Let $\gamma_1 \in \Pi(\mu^+_{\eta,\Delta}, \nu^+)$ and $\gamma_2 \in \Pi(\mu^-_{\eta,\Delta}, \nu^-)$ be optimal transport plans for the cost c_u, and let $\gamma_3 \in \Pi(\nu^+, \nu^-)$ be the optimal transport plan for the fairness penalty $c_{\mathcal{D}}$.

By the Gluing Lemma (e.g., Lemma 5.5 in Santambrogio [36]), we can construct a joint probability measure ρ over the product space $\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}$, with coordinates denoted (z_1, z_2, y_1, y_2), such that its respective two-dimensional marginals match γ_1, γ_2, and γ_3.

The global penalized objective $(P_{\lambda,\mathcal{D}})$ can then be evaluated as a single integral over this joint measure ρ

$$\int \left[c_u(z_1, y_1) + c_u(z_2, y_2) + \lambda c_{\mathcal{D}}(y_1, y_2)\right] \mathrm{d}\rho(z_1, z_2, y_1, y_2).$$

For every tuple (z_1, z_2, y_1, y_2) in the support of ρ, the integrand is bounded from below by its pointwize infimum over all possible target values $y_1, y_2 \in \mathbb{R}$ for the given pair (z_1, z_2). Substituting this pointwise minimum, $C_{\lambda,\mathcal{D}}(z_1, z_2)$, into the integral yields a lower bound. Furthermore, letting $\pi \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})$ denote the projection of ρ onto its first two variables (z_1, z_2), we obtain

$$\mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \nu^+) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \nu^-) + \lambda\, \mathrm{OT}_{c_{\mathcal{D}}}(\nu^+, \nu^-) \geq \int_{\mathbb{R}^2 \times \mathbb{R}^2} C_{\lambda,\mathcal{D}}(z_1, z_2)\, \mathrm{d}\pi(z_1, z_2).$$

Taking the infimum over all valid source couplings π establishes that $\min_\pi \int C_{\lambda,\mathcal{D}}\, \mathrm{d}\pi$ is a global lower bound for $(P_{\lambda,\mathcal{D}})$.

Existence and achieving the bound. The cost $C_{\lambda,\mathcal{D}}$ is lower semi-continuous as an infimum of lower semi-continuous costs. It is also clearly lower bounded since $c_{\mathcal{D}}$ is bounded from below. Therefore, we can apply Theorem 1.5 of Santambrogio [36], which gives us an optimal coupling $\pi^* \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})$. We denote the objective function

$$\Phi_{z_1,z_2}(y_1, y_2) := \frac{(h_1 - y_1)^2}{|d_1|} + \frac{(h_2 - y_2)^2}{|d_2|} + \lambda c_{\mathcal{D}}(y_1, y_2).$$

This function is lower semi-continuous and coercive, and therefore has a nonempty, compact set of minimizers. Since $\Phi_{z_1,z_2}(y_1, y_2)$ is measurable with respect to the variable (z_1, z_2), the Measurable Maximum Theorem (see Theorem 18.19 in Aliprantis and Border [3]) applies, which implies that there exists a measurable map $(y_1^*, y_2^*) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ such that $(y_1^*(z_1, z_2), y_2^*(z_1, z_2))$ is a minimizer of Φ_{z_1,z_2} for all $(z_1, z_2) \in \mathbb{R} \times \mathbb{R}$.

We can therefore push forward π^* via the map $(z_1, z_2) \mapsto (z_1, z_2, y_1^*, y_2^*)$ and construct a joint measure ρ^*. The marginals of ρ^* over its third and fouth coordinates (y_1 and y_2) define target distributions ν^{+*} and ν^{-*}. Because ρ^* constitutes a valid joint transport plan for these targets, the optimal transport cost is bounded above by the integral over ρ^*, yielding

$$\mathrm{OT}_{c_u}(\mu^+_{\eta,\Delta}, \nu^{+*}) + \mathrm{OT}_{c_u}(\mu^-_{\eta,\Delta}, \nu^{-*}) + \lambda\, \mathrm{OT}_{c_{\mathcal{D}}}(\nu^{+*}, \nu^{-*}) \leq \int_{\mathbb{R}^2 \times \mathbb{R}^2} C_{\lambda,\mathcal{D}}(z_1, z_2)\, \mathrm{d}\pi^*(z_1, z_2).$$

Combined with the previously established lower bound, this proves the equality (11), and that ν^{+*} and ν^{-*} are minimizers of $(P_{\lambda,\mathcal{D}})$.

Moreover, the previous argument clearly shows that for *any* optimal transport plan π^* and *any* choice of measurable minimizers (y_1^*, y_2^*), the measures ν^{+*} and ν^{-*} obtained through the previous construction are minimizers of $(P_{\lambda,\mathcal{D}})$, concluding the proof of the lemma. □

A.4 Proof of Prop. 1 (FAUST-U $\mathcal{W}_2$)

Proposition 1 (FAUST-U $\mathcal{W}_2$). *Under Assumption 1, Problem $(P_{\lambda,\mathcal{W}_2})$ is equivalent to the OT problem*

$$\pi^* \in \arg\min_{\pi \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})} \int_{\mathbb{R}^2 \times \mathbb{R}^2} C^u_{\lambda,c_2}(z_1, z_2)\, \mathrm{d}\pi(z_1, z_2), \tag{5}$$

where $C^u_{\lambda,c_2}(z_1,z_2) = \frac{\lambda}{1+\lambda(|d_1|+|d_2|)}(h_1-h_2)^2$ *for* $z_1=(h_1,d_1)$ *and* $z_2=(h_2,d_2)$. ***Moreover, a solution to the almost-fair regression problem in*** (3) ***is given by Eq.*** (4) ***with***

$$T^+_{\lambda,\mathcal{W}_2^2}(z_1) = \frac{(1+\lambda|d_2|)h_1+\lambda|d_1|h_2}{1+\lambda(|d_1|+|d_2|)} \quad \textit{and} \quad T^-_{\lambda,\mathcal{W}_2^2}(z_2) = \frac{\lambda|d_2|h_1+(1+\lambda|d_1|)h_2}{1+\lambda(|d_1|+|d_2|)} \tag{6}$$

for all pairs (z_1,z_2) *in the support of* π^*.

Proof. By Lemma 2, the global problem reduces to evaluating the pointwise relaxed cost for $\mathcal{D}(y_1,y_2) = (y_1-y_2)^2$. Fix $z_i=(h_i,d_i)$ with absolute density weights $a_i=|d_i|>0$, and minimize

$$\Phi_{z_1,z_2}(y_1,y_2) = \frac{(h_1-y_1)^2}{a_1} + \frac{(h_2-y_2)^2}{a_2} + \lambda(y_1-y_2)^2.$$

Because $\Phi := \Phi_{z_1,z_2}$ is strictly convex with respect to y_1 and y_2, the optimal pointwise targets are characterized by the first-order optimality conditions, which can be expressed as

$$\begin{cases} \frac{2(y_1-h_1)}{a_1} + 2\lambda(y_1-y_2) = & 0, \\ \frac{2(y_2-h_2)}{a_2} - 2\lambda(y_1-y_2) = & 0. \end{cases}$$

Solving the system yields the optimal pointwise targets y_1^* and y_2^*. Let $u := y_1^* - y_2^*$. From the optimality conditions we have

$$\begin{cases} y_1^* - h_1 = & -\lambda a_1 u, \\ y_2^* - h_2 = & \lambda a_2 u, \end{cases}$$

Subtracting these identities gives $u = (h_1-h_2) - \lambda(a_1+a_2)u$, which resolves to $u = \frac{h_1-h_2}{1+\lambda(a_1+a_2)}$. On the one hand, plugging this value into the first-order conditions yield

$$\begin{cases} y_1^* & = \frac{h_1(1+\lambda a_2)+\lambda h_2 a_1}{1+\lambda(a_1+a_2)}, \\ y_2^* & = \frac{h_2(1+\lambda a_1)+\lambda h_1 a_2}{1+\lambda(a_1+a_2)}. \end{cases} \tag{12}$$

On the other hand, plugging these back into Φ at the optimum yields

$$\begin{aligned} C_{\lambda,c_2}(z_1,z_2) =& \Phi(y_1^*,y_2^*) = \frac{(\lambda a_1 u)^2}{a_1} + \frac{(\lambda a_2 u)^2}{a_2} + \lambda u^2 \\ =& \lambda u^2 \left(1+\lambda(a_1+a_2)\right) \\ =& \lambda \frac{(h_1-h_2)^2}{(1+\lambda(a_1+a_2))^2}\left(1+\lambda(a_1+a_2)\right) \\ =& \frac{\lambda}{1+\lambda(a_1+a_2)}(h_1-h_2)^2. \end{aligned}$$

According to Lemma 2, a solution (ν^{+*},ν^{-*}) of $(P_{\lambda,\mathcal{D}})$ is given by the marginal distributions of $(y_1^*(Z_1,Z_2), y_2^*(Z_1,Z_2))$ where $(Z_1,Z_2)\sim\pi^*$ and π^* is as in (5). Our next goal is to show that under Assumption 1, the random variable $y_1^*(Z_1,Z_2)$ does not depend on Z_2. To do so, we closely follow the arguments of the Step 2 of Theorem 4 in [16]. The only difference between the two settings is that the cost $C_{\infty,c_2}(z_1,z_2) = (h_1-h_2)^2/(a_1+a_2)$ is replaced by the cost $C_{\lambda,c_2}(z_1,z_2)$. These two costs share the essential property that they are quadratic whenever a_1 and a_2 are fixed, which is enough to make the arguments of the proof in Divol and Gaucher [16] work.

Let us give some additional details. The proof is based on Kantorovich duality, see Section 2.1 in [16] for an introduction to this notion. There exists a function ψ, called a Kantorovich potential, that satisfies the following properties. Let $\phi: z_2 \mapsto \sup_{z_1\in\mathcal{Z}}(\psi(z_1) - C_{\lambda,c_2}(z_1,z_2))$ be the C_{λ,c_2}-transform of ψ. Then,

1. for all z_1, $\psi(z_1) = \inf_{z_2}(\phi(z_2) + C_{\lambda,c_2}(z_1,z_2))$;
2. π^* is supported on the C^u_{λ,c_2}-subdifferential of a Kantorovich potential ψ, that is the set

$$\Gamma = \{(z_1,z_2)\in\mathcal{Z}\times\mathcal{Z}:\ \psi(z_1)-\phi(z_2) = C_{\lambda,c_2}(z_1,z_2)\}.$$

The arguments used to prove Lemma 3 in Divol and Gaucher [16] can be reproduced to show that under Assumption 1, the Kantorovich potential ψ is differentiable in the direction h_1 for $\mu^+_{\eta,\Delta}$-almost every $z_1 = (h_1, d_1)$ (this follows from the fact that, when (d_1, d_2) is fixed, the cost $C_{\lambda,c_2}(z_1, z_2)$ is proportional to $(h_1 - h_2)^2$).

Let z_1 be a point in the support of $\mu^+_{\eta,\Delta}$ where $\partial_{h_1}\psi$ exists, and let z_2 be such that (z_1, z_2) is in the support of π^* (and therefore in Γ). The definitions of ϕ and Γ imply that the function $z \mapsto \psi(z) - C_{\lambda,c_2}(z, z_2)$ attains its maximum at z_1. Differentiating, we find that

$$\partial_{h_1}\psi(z_1) = \partial_{h_1} C^u_{\lambda,c_2}(z_1, z_2) = \frac{2\lambda(h_1 - h_2)}{1 + \lambda(a_1 + a_2)} = 2\lambda u.$$

Isolating $u = \frac{1}{2\lambda}\partial_{h_1}\psi(z_1)$ and substituting it into the optimality condition for y_1^* gives

$$y_1^*(z_1, z_2) = h_1 - \lambda a_1 u = h_1 - \frac{|d_1|}{2}\partial_{h_1}\psi(z_1).$$

In other words, the value $y_1^*(z_1, z_2)$ does not depend on the choice of z_2 with (z_1, z_2) in the support of π^* for $\mu^+_{\eta,\Delta}$-almost all z_1, which is what we wanted to show. A symmetric argument holds for y_2^*, concluding to the existence of deterministic maps $T^+_{\lambda,\mathcal{W}_2^2}$ and $T^-_{\lambda,\mathcal{W}_2^2}$ solving $(P_{\lambda,\mathcal{D}})$.

Together with Lemma 1 and Lemma 2, this proves Theorem 1 for the $\mathcal{W}_2^2$ unfairness penalty. Moreover, the proof of Prop. 1 is concluded by noticing that the expression of the optimal transport maps are given by Eq. (12). □

Let us note that Prop. 1 implies in particular the result of Theorem 1 in the case $c_{\mathcal{D}} = c_2$.

Remark A.1 (Asymptotic compatibility with exact fairness). As a sanity check, let us verify that the solution of the problem $(P_{\lambda,\mathcal{D}})$ converge to the solution of the fair regression problem (P_∞) as $\lambda \to \infty$.

Consider the relaxed objective $(P_{\lambda,\mathcal{D}})$ and its reduced form Eq. (11) with the pointwise cost $C^u_{\lambda,c_2}(z_1, z_2)$. Let $z_i = (h_i, d_i)$ with $a_i = |d_i| > 0$. The inner minimization admits the explicit map evaluations given in Eq. (6), the gap between the predictions strictly vanishes

$$T^+_{\lambda,\mathcal{W}_2^2}(z_1) - T^-_{\lambda,\mathcal{W}_2^2}(z_2) = \frac{h_1 - h_2}{1 + \lambda(a_1 + a_2)} \xrightarrow[\lambda\to\infty]{} 0, \tag{13}$$

and moreover, both maps naturally converge to the weighted barycenter

$$T^+_{\lambda,\mathcal{W}_2^2}(z_1),\ T^-_{\lambda,\mathcal{W}_2^2}(z_2) \xrightarrow[\lambda\to\infty]{} \frac{a_2 h_1 + a_1 h_2}{a_1 + a_2} =: y^*(z_1, z_2), \tag{14}$$

where $y^*(z_1, z_2)$ is the minimizer of the unrelaxed two-to-one cost $\inf_y\{c_u(z_1, y) + c_u(z_2, y)\}$.

At the level of pairwise costs, the closed form expression of the cost yields the pointwise convergence

$$C^u_{\lambda,c_2}(z_1, z_2) = \frac{\lambda}{1 + \lambda(a_1 + a_2)}(h_1 - h_2)^2 \xrightarrow[\lambda\to\infty]{} \frac{(h_1 - h_2)^2}{a_1 + a_2} =: C_\infty(z_1, z_2),$$

which is the unrelaxed OT cost.

Let now $\pi^*_\lambda \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})$ be an optimal coupling for the reduced problem in Eq. (11). Denote by V_λ the optimal value in Eq. (11). Then V_λ is nondecreasing in λ and satisfies $V_\lambda \le V_\infty$, where V_∞ is the optimal value of the unrelaxed problem (since $C^u_{\lambda,c_2} \le C_\infty$). We claim that we actually have

$$V_\lambda \xrightarrow[\lambda\to\infty]{} V_\infty.$$

Indeed, if were not the case, we could find a subsequence (λ_k) such that V_λ converges to some smaller limit along this subsequence. As the set $\Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})$ is compact for the weak convergence [39, Lemma 4.4], we may find a transport plan $\pi^*_\infty \in \Pi(\mu^+_{\eta,\Delta}, \mu^-_{\eta,\Delta})$ such that a subsequence of $\pi^*_{\lambda_k}$ converges weakly towards π^*_∞. Let us still denote this subsequence by $\pi^*_{\lambda_k}$ for the sake of simplicity. We can then apply Fatou's lemma [20] to find that

$$V_\infty > \liminf_{k\to\infty} V_{\lambda_k} = \liminf_{k\to\infty} \int C^u_{\lambda_k,c_2}\,\mathrm{d}\pi^*_{\lambda_k} \ge \int C_\infty\,\mathrm{d}\pi^*_\infty \ge V_\infty,$$

which yields a contradiction.

Finally, define the associated optimal outputs by

$$Y_{1,\lambda} := T^{+}_{\lambda,\mathcal{W}_2^2}(Z_1), \qquad Y_{2,\lambda} := T^{-}_{\lambda,\mathcal{W}_2^2}(Z_2), \qquad (Z_1, Z_2) \sim \pi^*_\lambda,$$

so that $\nu^*_{1,\lambda} = \mathcal{L}(Y_{1,\lambda}) = (T^{+}_{\lambda,\mathcal{W}_2^2})_{\#}\mu^{+}_{\eta,\Delta}$ and $\nu^*_{2,\lambda} = \mathcal{L}(Y_{2,\lambda}) = (T^{-}_{\lambda,\mathcal{W}_2^2})_{\#}\mu^{-}_{\eta,\Delta}$. Consider moreover the coupling between these two output laws induced by the same construction,

$$\gamma_\lambda := (T^{+}_{\lambda,\mathcal{W}_2^2}, T^{-}_{\lambda,\mathcal{W}_2^2})_{\#}\pi^*_\lambda \in \Pi(\nu^*_{1,\lambda}, \nu^*_{2,\lambda}).$$

By definition of $\mathcal{W}_2$, we have

$$\begin{aligned}
\mathcal{W}_2^2(\nu^*_{1,\lambda}, \nu^*_{2,\lambda}) \;&\le\; \int_{\mathbb{R}^2} (y_1 - y_2)^2 \, d\gamma_\lambda(y_1, y_2) \;=\; \mathbb{E}_{\pi^*_\lambda}\big[(Y_{1,\lambda} - Y_{2,\lambda})^2\big] \\
&= \int \frac{(h_1 - h_2)^2}{(1 + \lambda(|d_1| + |d_2|))^2} \, \mathrm{d}\pi^*_\lambda(z_1, z_2) \\
&\le \frac{2}{\lambda} \left(\int \frac{h_1^2}{|d_1|} \, \mathrm{d}\mu^{+}_{\eta,\Delta}(z_1) + \int \frac{h_1^2}{|d_1|} \, \mathrm{d}\mu^{-}_{\eta,\Delta} \right)(z_2),
\end{aligned}$$

where we use the inequality $(h_1 - h_2)^2 \le 2h_1^2 + 2h_2^2$. Since $\int \frac{h^2}{|d|} \, \mathrm{d}\mu^{\pm}_{\eta,\Delta}(z) = \int_{\mathcal{X}_\pm} \eta(x)^2 \, \mathrm{d}\chi(x) \le \mathbb{E}[\eta(X)^2] < \infty$, we have $\mathcal{W}_2^2(\nu^*_{1,\lambda}, \nu^*_{2,\lambda}) \to 0$ as $\lambda \to \infty$. In particular, the two output laws become arbitrarily close in $\mathcal{W}_2$.

Therefore the relaxed solutions asymptotically enforce a single common output distribution; if the unrelaxed barycenter is unique, this common limit must coincide with the unrelaxed barycenter law.

A.5 Proof of Prop. 2 (FAUST-U TV)

Proposition 2 (FAUST-U TV). *Under Assumption 1, Problem $(P_{\lambda,\,\mathrm{TV}})$ is equivalent to the OT problem*

$$\pi^* \in \arg \min_{\pi \in \Pi(\mu^{+}_{\eta,\Delta}, \mu^{-}_{\eta,\Delta})} \int_{\mathbb{R}^2 \times \mathbb{R}^2} C^u_{\lambda,c_0}(z_1, z_2) \, \mathrm{d}\pi(z_1, z_2), \tag{7}$$

where $C^u_{\lambda,c_0}(z_1, z_2) = \min\left(\lambda, \; \frac{(h_1 - h_2)^2}{|d_1| + |d_2|}\right)$. Moreover, a solution to the fair regression problem in Eq. (3) *is given by Eq.* (4) *with*

$$\begin{cases}
T^{+}_{\lambda,\,\mathrm{TV}}(z_1) = T^{-}_{\lambda,\,\mathrm{TV}}(z_2) = \dfrac{|d_2| h_1 + |d_1| h_2}{|d_1| + |d_2|} & \text{if} \quad \frac{(h_1 - h_2)^2}{|d_1| + |d_2|} \le \lambda, \\[2ex]
T^{+}_{\lambda,\,\mathrm{TV}}(z_1) = h_1, \quad T^{-}_{\lambda,\,\mathrm{TV}}(z_2) = h_2 & \text{if} \quad \frac{(h_1 - h_2)^2}{|d_1| + |d_2|} > \lambda,
\end{cases} \tag{8}$$

for all pairs $z_1 = (h_1, d_1)$ and $z_2 = (h_2, d_2)$ in the support of π^*.

Proof of Prop. 2. By Lemma 2, the global unaware problem reduces to evaluating the pointwise relaxed cost $C^u_{\lambda,c_0}(z_1, z_2)$. For any fixed pair $z_1 = (h_1, d_1)$ and $z_2 = (h_2, d_2)$ with absolute density weights $a_1 = |d_1| > 0$ and $a_2 = |d_2| > 0$, we seek to minimize the objective

$$\Phi(y_1, y_2) = \frac{(h_1 - y_1)^2}{a_1} + \frac{(h_2 - y_2)^2}{a_2} + \lambda \mathbb{1}_{\{y_1 \neq y_2\}}.$$

We divide the pointwise optimization into two mutually exclusive cases depending on the indicator function.

Case 1: We restrict the domain to $y_1 \neq y_2$.

The indicator evaluates to 1, yielding the objective $\Phi_{y_1 \neq y_2}(y_1, y_2) = \frac{(h_1 - y_1)^2}{a_1} + \frac{(h_2 - y_2)^2}{a_2} + \lambda$. Because $a_1, a_2 > 0$, the unconstrained minimum of this convex quadratic is trivially achieved at $y_1^* = h_1$ and $y_2^* = h_2$. Assuming $h_1 \neq h_2$, this solution respects the domain restriction, and the minimum cost for this branch is exactly λ.

Case 2: We restrict the domain to $y_1 = y_2 = y$.
The indicator evaluates to 0, simplifying the objective to a single-variable quadratic $\Phi_{y_1=y_2}(y) = \frac{(h_1-y)^2}{a_1} + \frac{(h_2-y)^2}{a_2}$. The first-order optimality condition is $-\frac{2(h_1-y)}{a_1} - \frac{2(h_2-y)}{a_2} = 0$. Solving for y yields the weighted barycenter

$$y^* = \frac{(1/a_1)h_1 + (1/a_2)h_2}{1/a_1 + 1/a_2} = \frac{a_2h_1 + a_1h_2}{a_1 + a_2}. \tag{15}$$

We now plug y^* back into the objective

$$\Phi(y^*, y^*) = \frac{1}{a_1}\left(h_1 - \frac{a_2h_1 + a_1h_2}{a_1 + a_2}\right)^2 + \frac{1}{a_2}\left(h_2 - \frac{a_2h_1 + a_1h_2}{a_1 + a_2}\right)^2.$$

Simplifying the terms inside the squares gives $\frac{a_1(h_1-h_2)}{a_1+a_2}$ for the first and $\frac{-a_2(h_1-h_2)}{a_1+a_2}$ for the second. Squaring these and multiplying by their respective outer weights yields

$$\begin{aligned}\Phi(y^*, y^*) =& \frac{1}{a_1}\frac{a_1^2(h_1-h_2)^2}{(a_1+a_2)^2} + \frac{1}{a_2}\frac{a_2^2(h_1-h_2)^2}{(a_1+a_2)^2} \\ =& \frac{a_1+a_2}{(a_1+a_2)^2}(h_1-h_2)^2 \\ =& \frac{(h_1-h_2)^2}{a_1+a_2}.\end{aligned}$$

Conclusion. According to Lemma 2, a solution (ν^{+*}, ν^{-*}) of $(P_{\lambda,\mathcal{D}})$ is given by the marginals distributions of $(y_1^*(Z_1, Z_2), y_2^*(Z_1, Z_2))$ where $(Z_1, Z_2) \sim \pi^*$ and π^* is as in (7). Precisely, the map (y_1^*, y_2^*) is defined by letting $(y_1^*(z_1, z_2), y_2^*(z_1, z_2)) = (h_1, h_2)$ if $C^u_{\lambda,c_0}(z_1, z_2) = \lambda$, and $(y_1^*(z_1, z_2), y_2^*(z_1, z_2)) = (y^*, y^*)$ with y^* defined in (15) otherwise.

Our next goal is to show that under Assumption 1, the random variable $y_1^*(Z_1, Z_2)$ does not depend on Z_2. We use once again Kantorovich duality.

Recall that there exists a function ψ, called a Kantorovich potential, that satisfies the following properties. Let $\phi : z_2 \mapsto \sup_{z_1 \in \mathcal{Z}}(\psi(z_1) - C_{\lambda,c_0}(z_1, z_2))$ be the C_{λ,c_0}-transform of ψ. Then, π^* is supported on the C^u_{λ,c_0}-subdifferential of a Kantorovich potential ψ, that is the set

$$\Gamma = \{(z_1, z_2) \in \mathcal{Z} \times \mathcal{Z} :\ \psi(z_1) - \phi(z_2) = C_{\lambda,c_0}(z_1, z_2)\}.$$

Furthermore, the function ψ is C_{λ,c_0}-concave, meaning that for all z_1, $\psi(z_1) = \inf_{z_2}(\phi(z_2) + C_{\lambda,c_0}(z_1, z_2))$. Using the definition of C_{λ,c_0}, we find that

$$\psi(z_1) = \inf_{z_2}(\phi(z_2) + \min(C_\infty(z_1, z_2), \lambda)) = \min\left(\tilde{\psi}(z_1), \inf_{z_2}\phi(z_2) + \lambda\right)$$

where $\tilde{\psi}(z_1) = \inf_{z_2}(\phi(z_2) + C_\infty(z_1, z_2))$ defines a C_∞-concave function for $C_\infty(z_1, z_2) = \frac{(h_1-h_2)^2}{a_1+a_2}$. According to Lemma 3 in [16], under Assumption 1, the function $\tilde{\psi}$ is $\mu^+_{\eta,\Delta}$-almost everywhere differentiable at $z_1 = (h_1, d_1)$ in the direction h_1.

Let $\mathcal{Z}_< = \{z_1 \in \mathcal{Z} :\ \psi(z_1) = \tilde{\psi}(z_1)\}$. If $(z_1, z_2) \in \Gamma$ with $z_1 \in \mathcal{Z}_<$, then

$$\begin{aligned}C_\infty(z_1, z_2) \geq \min(\lambda, C_\infty(z_1, z_2)) = C_{\lambda,c_0}(z_1, z_2) = \psi(z_1) - \phi(z_2) \\ = \tilde{\psi}(z_1) - \phi(z_2) \geq C_\infty(z_1, z_2).\end{aligned}$$

Thus, $\psi(z_1) - \phi(z_2) = C_\infty(z_1, z_2)$. In other words, the set $\{(z_1, z_2) \in \Gamma :\ z_1 \in \mathcal{Z}_<\}$ is included in the C_∞-subdifferential of $\tilde{\psi}$. It is proven in [16] (as explained in the previous section) that if z_1 is a point where $\tilde{\psi}$ is differentiable at z_1 in the direction h_1, and if (z_1, z_2) is in the C_∞-subdifferential of $\tilde{\psi}$, then the quantity

$$y_1^*(z_1, z_2) = \frac{a_2h_1 + a_1h_2}{a_1 + a_2}$$

does not depend on the choice of z_2. Thus, the map $T^+_{\lambda,\,\mathrm{TV}}(z_1) = \frac{a_2 h_1 + a_1 h_2}{a_1 + a_2}$ is well-defined for $\mu^+_{\eta,\Delta}$-almost all $z_1 \in \mathcal{Z}_<$.

If $z_1 \in \mathcal{Z}_= = \mathcal{Z}\backslash\mathcal{Z}_<$, then $\tilde{\psi}(z_1) > \inf_{z_2} \phi(z_2) + \lambda$. Thus, if $(z_1, z_2) \in \Gamma$, then

$$\begin{aligned}\lambda &\geq \min(\lambda, C_\infty(z_1, z_2)) = C_{\lambda,c_0}(z_1, z_2) = \psi(z_1) - \phi(z_2)\\ &= \inf_{z_2'} \phi(z_2') + \lambda - \phi(z_2) \geq \lambda.\end{aligned}$$

Thus $C_{\lambda,c_0}(z_1, z_2) = \lambda$ and thus $y_1^*(z_1, z_2) = h_1$ clearly only depends on z_1. This shows that the map $T^+_{\lambda,\,\mathrm{TV}}$ is well-defined $\mu^+_{\eta,\Delta}$-almost everywhere.

Similarly, we show that the map $T^-_{\lambda,\,\mathrm{TV}}$ is well-defined $\mu^-_{\eta,\Delta}$-almost everywhere.

We then conclude using Lemma 1 to show that solutions of Equation (3) take the form described in the lemma. □

Let us note that Prop. 1 implies in particular the result of Theorem 1 in the case $c_{\mathcal{D}} = c_0$.

B Proofs for the relaxed aware setting

As discussed in Section 2.1, the aware setting corresponds to the boundary case of Problem $(P_{\lambda,\mathcal{D}})$, where the sensitive attribute is observed, that is we observe (X, S) instead of X. In that case, the Jordan decomposition partitions the space into the demographic groups, yielding uniform absolute density weights $a_1 = 1/p^+$ for group $S = +$ and $a_2 = 1/p^-$ for group $S = -$. Then, χ^+ and χ^- have separate supports, with $\Delta(x, +) = 1/p^+$ for $(x, +)$ in the support of χ^+, $\Delta(x, -) = -1/p^-$ for $(x, -)$ in the support of χ^-. The measure $\mu^+_{\eta,\Delta}$ is simply the law of $(\eta(X, +), 1/p^+)$ where $X \sim \chi^+$, and $\mu^-_{\eta,\Delta}$ is the law of $(\eta(X, -), -1/p^-)$ where $X \sim \chi^-$.

The two-dimensional costs described in Prop. 1 and Prop. 2 now depend effectively only on (h_1, h_2), that is these costs become one-dimensional. In that case, the solution of the corresponding optimal transport problem can be formulated in a simpler fashion. Let us describe it more precisely.

Exact Aware Solution. First, in the limit $\lambda \to \infty$ (no penalization), as established by Chzhen et al. [12], the exact Wasserstein barycenter for the 1D aware setting admits a closed-form. The fair prediction $f^*_\infty(x, s)$ for an individual is obtained by evaluating their relative rank within their own demographic group via the cumulative distribution function $F_{\mu^s_\eta}$, and mapping it to the weighted quantiles of both groups

$$f^*_\infty(x, s) = q^* \circ F_{\mu^s_\eta}(\eta(x, s)) := \left(p^+ Q_{\mu^+_\eta} + p^- Q_{\mu^-_\eta}\right) \circ F_{\mu^s_\eta}(\eta(x, s))$$

where $q^* : t \in (0, 1) \mapsto p^+ Q_{\mu^+_\eta}(t) + p^- Q_{\mu^-_\eta}(t)$ is the barycenter map.

B.1 Aware $\mathcal{W}_2$ relaxation

Here we consider the fairness penalty $\mathcal{D}(\cdot, \cdot) = OT_{c_2}(\cdot, \cdot)$ for $c_2(y, y') = (y - y')^2$.

Corollary 1 (FAUST-A $\mathcal{W}_2$). *Under Assumption 1, in the awareness setting S, the solution to* (3) *is given by the optimal predictor $f^*_{\lambda,\mathcal{W}_2}(x, s)$, which interpolates between the unconstrained predictor $\eta(x, s)$ and the exact fair predictor $f^*_\infty(x, s)$*

$$f^*_{\lambda,\mathcal{W}_2}(x, s) = (1 - \alpha) f^*_\infty(x, s) + \alpha\eta(x, s), \quad \textit{where} \quad \alpha = \frac{p^+ p^-}{p^+ p^- + \lambda}.$$

This explicitly recovers the closed-form geodesic interpolation derived by Chzhen and Schreuder [11].

Proof. In the framework of Chzhen and Schreuder [11], the fairness penalty is defined as the weighted variance $p^+ p^- \mathcal{W}_2^2(\nu^+, \nu^-)$. Thus, our λ-penalized objective is equivalent to their formulation under the effective penalty $\lambda_C = \lambda/(p^+ p^-)$.

According to their Proposition 4.7, this problem admits a relative improvement governed by an interpolation parameter $\rho = (1 + \lambda_C)^{-2}$. Following their Proposition 4.1, the optimal relaxed predictor is uniquely defined as the convex combination of the perfectly fair projection and the unconstrained base predictor, weighted by $1 - \sqrt{\rho}$ and $\sqrt{\rho}$ respectively. Substituting our penalty yields exactly

$$\sqrt{\rho} = \left(1 + \frac{\lambda}{p^+p^-}\right)^{-1} = \frac{p^+p^-}{p^+p^- + \lambda} = \alpha.$$

We now show that our generalized unaware reduction perfectly recovers this geometric interpolation from the maps defined in Prop. 1. Substituting the deterministic aware weights $a_1 = 1/p^+$ and $a_2 = 1/p^-$, the interpolation denominator becomes

$$1 + \lambda(a_1 + a_2) = 1 + \lambda\left(\frac{p^+ + p^-}{p^+p^-}\right),$$

and because $p^+ + p^- = 1$, this simplifies to

$$1 + \lambda(a_1 + a_2) = 1 + \frac{\lambda}{p^+p^-} = \alpha^{-1}.$$

Thus, $C^u_{\lambda,c_2}(z_1, z_2) = \alpha\lambda(h_1 - h_2)^2$ for $z_1 = (h_1, 1/p^+)$ and $z_2 = (h_2, -1/p^-)$.

The optimal transport problem minimizing this cost is the monotone matching, obtained by letting $h_1 = Q_{\mu^+_\eta}(t)$ and $h_2 = Q_{\mu^-_\eta}(t)$ for a given $t \in (0, 1)$.

We then find from Eq. (6)

$$\begin{aligned} T^+_{\lambda,\mathcal{W}_2^2}(z_1) =& \alpha\left[\left(1 + \frac{\lambda}{p^-}\right) Q_{\mu^+_\eta}(t) + \frac{\lambda}{p^+} Q_{\mu^-_\eta}(t)\right] \\ =& \alpha Q_{\mu^+_\eta}(t) + \alpha\lambda\left(\frac{Q_{\mu^+_\eta}(t)}{p^-} + \frac{Q_{\mu^-_\eta}(t)}{p^+}\right). \end{aligned}$$

Recognizing from the definition of α that $\alpha\lambda = (1 - \alpha)p^+p^-$, the second term simplifies to $(1-\alpha)(p^+Q_{\mu^+_\eta}(t)+p^-Q_{\mu^-_\eta}(t))$, which corresponds to $(1-\alpha)q^*(t)$, where q^* is the strict Wasserstein barycenter.

Evaluating this optimal quantile mapping at the individual's relative rank $t_{x,+} = F_{\mu^+_\eta}(\eta(x, +))$ recovers the predictor formulation for the positive group

$$f^*_{\lambda,\mathcal{W}_2}(x, +) = (1 - \alpha)f^*_\infty(x, +) + \alpha\eta(x, +),$$

and symmetrically for group $S = -$ using T^-. Thus, the unaware mapping retrieves the aware geodesic interpolation. □

B.2 Aware TV relaxation

Here we consider the fairness penalty $\mathcal{D}(\cdot, \cdot) = OT_{c_0}(\cdot, \cdot)$ for $c_0(y, y') = \mathbb{1}_{\{y \neq y'\}}$.

Corollary 2 (FAUST-A TV). *Under Assumption 1, in the awareness setting $S \in \{+, -\}$, the solution to* (3) *is characterized by the optimal transport plan $\pi^* \in \Pi(\mu^+_\eta, \mu^-_\eta)$ that minimizes the cost*

$$\pi^* \in \arg\min_{\pi \in \Pi(\mu^+_\eta, \mu^-_\eta)} \int_{\mathbb{R}\times\mathbb{R}} \min\left(\lambda,\ p^+p^-(h_1 - h_2)^2\right) \mathrm{d}\pi(h_1, h_2). \tag{16}$$

For any pair (h_1, h_2) in the support of π^*, the solution to Eq. (3) *is given by*

$$\begin{cases} f^*_{\lambda,\,\mathrm{TV}}(h_1, +) = f^*_{\lambda,\,\mathrm{TV}}(h_2, -) = p^+h_1 + p^-h_2 & \text{if} \quad p^+p^-(h_1 - h_2)^2 \leq \lambda, \\ f^*_{\lambda,\,\mathrm{TV}}(h_1, +) = h_1, \quad f^*_{\lambda,\,\mathrm{TV}}(h_2, -) = h_2 & \text{if} \quad p^+p^-(h_1 - h_2)^2 > \lambda. \end{cases}$$

Proof of Cor. 2. In the aware setting, the absolute density weights are constant and strictly determined by the group priors: $a_1 = |d_1| = \frac{1}{p^+}$ for the positive group and $a_2 = |d_2| = \frac{1}{p^-}$ for the negative group. We derive the aware solution by substituting these values directly into the results of Prop. 2.

First, we evaluate the denominator present in both the cost and the barycenter formulations

$$a_1 + a_2 = \frac{1}{p^+} + \frac{1}{p^-} = \frac{1}{p^+ p^-},$$

where we used that $p^+ + p^- = 1$. Substituting this into the threshold from Prop. 2 yields

$$\frac{(h_1 - h_2)^2}{a_1 + a_2} = p^+ p^- (h_1 - h_2)^2.$$

This recovers the threshold $p^+p^-(h_1 - h_2)^2 \leq \lambda$ and the truncated cost $\min\left(\lambda,\ p^+p^-(h_1 - h_2)^2\right)$ stated in the corollary.

Next, we substitute the weights into the optimal weighted barycenter y^* defined in Prop. 2

$$y^* = \frac{a_2 h_1 + a_1 h_2}{a_1 + a_2} = \left(\frac{1}{p^-} h_1 + \frac{1}{p^+} h_2\right) p^+ p^- = p^+ h_1 + p^- h_2.$$

Applying this simplified cost threshold and this simplified barycenter directly into the piecewise formulation of the deterministic maps from Prop. 2 exactly recovers the closed-form predictions for $f^*_{\lambda,\,\mathrm{TV}}$, concluding the proof.

Finally, we reconstruct the optimal map. According to Prop. 2, when the cost exceeds the threshold λ, the map outputs the unconstrained base prediction h_s. When the cost is below the threshold λ, the maps output the shared barycenter, recovering the formula in the statement of the corollary. □

C Unfairness of plug-in approaches

Replacing the unknown senstitive attribute S by a proxy $\hat{S}$ is a common practice [40]. A line of work [8, 27, 25, 5] has therefor studied the relationship between fairness with respect to a proxy sensitive attribute and fairness with respect to the true sensitive attribute. With the exception of [27], which establishes impossibility results, these works generally aim to show that, either under suitable assumptions on the proxy or empirically, low proxy-based unfairness translates into low unfairness with respect to the true sensitive attribute. These assumptions are often too strong to hold in practice, as illustrated by the unavoidable level of unfairness observed for the plug-in approach in our simulations.

Below, we state a result that may be considered folklore. We were unable to find a precise reference for it, and therefore include it together with a proof based on counterexamples. These counterexamples show that, in general, fairness with respect to a proxy does not imply fairness with respect to the true sensitive attribute.

Proposition 3. *Let $\hat{S}(X) = \arg max_{s\in\{+,-\}} \mathbb{P}(S = s|X)$ be the maximum a posteriori estimator of the sensitive attribute (ties broken arbitrarily). Then, for $f : \mathcal{X} \mapsto \mathbb{R}$,*

$$f(X) \perp\!\!\!\perp \hat{S}(X) \not\Rightarrow f(X) \perp\!\!\!\perp S.$$

Moreover, for $f : \mathcal{X} \times S \mapsto \mathbb{R}$,

$$f(X, S) \perp\!\!\!\perp S \not\Rightarrow f(X, \hat{S}(X)) \perp\!\!\!\perp S.$$

Prop. 3 proves that plug-in approaches, either enforcing DP with regards to $\hat{S}$ or replacing S by $\hat{S}$ in aware predictors generally fail to enforce demographic parity.

Proof. We prove Prop. 3 by providing simple counter-examples. (**First claim**) Consider the following distributions : $S \sim$ Rademacher, $X|S = + \sim 0.8 \times \mathcal{U}([0, 0.5]) + 0.2 \times \mathcal{U}([0.75, 1])$, $X|S = - \sim 0.6 \times \mathcal{U}([0.5, 0.75]) + 0.4 \times \mathcal{U}([0.75, 1])$, and $Y = X$. Then, $\hat{S}(x) = +$ if $X \leq 0.5$, and $\hat{S}(x) = -$ if $x > 0.5$. The optimal prediction function that is fair with regards to $\hat{S}(X)$ (i.e., the prediction function enforcing $f(X) \perp\!\!\!\perp \hat{S}(X)$ with minimal squared risk) is given by the maps to the Wasserstein

barycenter of the distributions $Y|\hat{S}(X) = + \sim \mathcal{U}([0, 0.5])$ and $Y|\hat{S}(X) = - \sim \mathcal{U}([0.5, 1])$ with weights $\mathbb{P}(\hat{S}(X) = +) = 0.4$ and $\mathbb{P}(\hat{S}(X) = -) = 0.6$ [12]. Its distribution is therefore $\mathcal{U}([0.3, 0.8])$, and the optimal fair prediction is

$$f(x) = \begin{cases} x + 0.3 & \text{if} \quad x \leq 0.5 \\ x - 0.2 & \text{else.} \end{cases}$$

By construction, $f(X) \perp\!\!\!\perp \hat{S}(X)$. However, it is easy to see that

$$\begin{aligned} \mathbb{P}(f(X) \in [0.3, 0.55]|S = +) &= \mathbb{P}(f(X) \in [0.3, 0.55], \hat{S}(X) = +|S = +) \\ &= \mathbb{P}(\hat{S}(X) = +|S = +)\mathbb{P}(f(X) \in [0.3, 0.55]|S = +, X \leq 0.5) \\ &= 0.8\mathbb{P}(X + 0.3 \in [0.3, 0.55]|S = +, X \leq 0.5) \\ &= 0.4. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbb{P}(f(X) \in [0.3, 0.55]|S = -) &= \mathbb{P}(X - 0.2 \in [0.3, 0.55]|S = -) \\ &= \mathbb{P}(X \in [0.5, 0.75]|S = -) \\ &= 0.6. \end{aligned}$$

Thus, $f(X) \not\perp\!\!\!\perp S$, which concludes the proof.

(**Second claim**) Consider the following distribution : $S \sim$ Rademacher, $X|S = + \sim \mathcal{U}([0, 1])$, $X|S = - \sim \mathcal{U}([0.5, 1])$, and $Y = X$. Then, $\hat{S}(x) = +$ if $X \leq 0.5$, and $\hat{S}(x) = -$ if $x > 0.5$. The optimal fair prediction function in the awareness setting is given by the maps to the Wasserstein barycenter, which is $\mathcal{U}([0.25, 1])$, and the optimal fair prediction is

$$f(x, s) = \begin{cases} 0.75x + 0.25 & \text{if} \quad s = + \\ 1.5x - 0.5 & \text{else.} \end{cases}$$

By construction, we have $f(X, S) \perp\!\!\!\perp S$. Yet, on the one hand,

$$\begin{aligned} \mathbb{P}(f(X, \hat{S}(X)) \in [0.625, 1]|S = +) &= \mathbb{P}(f(X, -) \in [0.625, 1]|S = +) \\ &= \mathbb{P}(1.5X - 0.5 \in [0.625, 1]|S = +) \\ &= \mathbb{P}(X \in [0.75, 1]|S = +) \\ &= 0.25. \end{aligned}$$

On the other hand,

$$\begin{aligned} \mathbb{P}(f(X, \hat{S}(X)) \in [0.625, 1]|S = -) &= \mathbb{P}(X \in [0.75, 1]|S = -) \\ &= 0.5. \end{aligned}$$

Thus, $f(X, \hat{S}(X)) \not\perp\!\!\!\perp S$. □

D Algorithm pseudocode

We provide the pseudocode of our algorithm given in Algorithm 1, as well as a justification of the reweighting of the empirical measures.

Justification of the reweighting step Recall that $(X_i)_{i \leq n} \overset{i.i.d.}{\sim} \chi$ for $\chi = p_+\chi_+ + p_-\chi_-$. To estimate the distributions $\mu^+_{\eta,\Delta} = (\eta(\cdot), \Delta(\cdot))_{\#} (\chi^+ - \chi^-)_+$ and $\mu^-_{\eta,\Delta} = (\eta(\cdot), \Delta(\cdot))_{\#} (\chi^+ - \chi^-)_-$, we note that for x such that $\Delta(x) > 0$, we have by Bayes rule $\mathbb{P}(S = s|X = x)\, \mathrm{d}\chi(x) = p_s\, \mathrm{d}\chi^s(x)$

$$\begin{aligned} \Delta(x)\, \mathrm{d}\chi(x) &= \left(\frac{\mathbb{P}(S = +|X = x)}{p_+} - \frac{\mathbb{P}(S = -|X = x)}{p_-} \right) \mathrm{d}\chi(x) \\ &\propto \mathrm{d}\left(\chi^+ - \chi^-\right)_+ (x), \end{aligned}$$

where the proportionality comes from defining $(\chi^+ - \chi^-)_+$ and $(\chi^+ - \chi^-)_-$ as the *normalized* Jordan decomposition of $(\chi^+ - \chi^-)$. Thus, the weighted empirical measure $(X_i)_{i \leq n}$ with weights

Algorithm 1 Unaware Fair Regression

Input: Base regressor $\hat{\eta}(x)$, sensitive attribute's classifier $\hat{p}(s|x)$, priors p^+ and p^-, dataset $\{x_k\}_{k=1}^n$
Output: Fair continuous map $\hat{f}_{fair}$

Step 1: Estimation and Jordan decomposition
For all $k \in \{1, \dots, n\}$, compute unconstrained predictions $\hat{\eta}_k \leftarrow \hat{\eta}(x_k)$.
For all $k \in \{1, \dots, n\}$, compute density ratio differences: $\hat{\Delta}_k \leftarrow \frac{\hat{p}(+|x_k)}{p^+} - \frac{\hat{p}(-|x_k)}{p^-}$
Define index sets: $\mathcal{I}^+ = \{i : \hat{\Delta}_i > \tau\}$ and $\mathcal{I}^- = \{j : \hat{\Delta}_j < -\tau\}$.
Compute normalized probability weights a_i (for $i \in \mathcal{I}^+$) and b_j (for $j \in \mathcal{I}^-$) so $\sum a_i = \sum b_j = 1$.

Step 2: Relaxed Optimal Transport
Cost $\mathbf{C}_{i,j} \leftarrow C_{\lambda,\mathcal{D}}\left((\hat{\eta}_i, \hat{\Delta}_i), (\hat{\eta}_j, \hat{\Delta}_j)\right)$ *(e.g., via Prop. 1 or Prop. 2)*
Pairwise targets $m^+_{i,j} \leftarrow T^+_{\lambda,\mathcal{D}}\left((\hat{\eta}_i, \hat{\Delta}_i), (\hat{\eta}_j, \hat{\Delta}_j)\right)$ and $m^-_{i,j} \leftarrow T^-_{\lambda,\mathcal{D}}\left((\hat{\eta}_i, \hat{\Delta}_i), (\hat{\eta}_j, \hat{\Delta}_j)\right)$
Solve discrete OT plan $\mathbf{\Pi}$ between $\hat{\mu}^+$(weights a_i) and $\hat{\mu}^-$(weights b_j) using cost matrix $\mathbf{C}$.

Step 3: Construct the fair targets
Initialize pseudo-label vector $f = (f_1, \dots, f_n)$.
for $k = 1$ to n **do**
  if $k \in \mathcal{I}^+$ **then**
    $f_k \leftarrow \frac{1}{a_k} \sum_{j \in \mathcal{I}^-} \mathbf{\Pi}_{k,j} m^+_{k,j}$
  else if $k \in \mathcal{I}^-$ **then**
    $f_k \leftarrow \frac{1}{b_k} \sum_{i \in \mathcal{I}^+} \mathbf{\Pi}_{i,k} m^-_{i,k}$
  else
    $f_k \leftarrow \hat{\eta}_k$
  end if
end for

Step 4: Fit final regressor
Train the continuous map $\hat{f}_{fair}$ on inputs $\{(\hat{\eta}_k, \hat{\Delta}_k)\}$ to predict targets $\{f_k\}$.

$\Delta(X_i)_+$ is a empirical (unbiased) version of the measure $(\chi^+ - \chi^-)_+$, so $(\hat{\eta}_j, \hat{d}_j)_{j \in \mathcal{I}^+}$ with corresponding weights $(\hat{d}_i)_{i \in \mathcal{I}^+}$ is an estimate for the measure $\mu^+_{\eta,\Delta}$.

Similarly, for x such that $\Delta(x) < 0$, we have $|\Delta(x)|\,\mathrm{d}\chi(x) \propto \mathrm{d}\,(\chi^+ - \chi^-)_-\,(x)$, and the weighted empirical measure $(X_i)_{i \leq n}$ with weights $-\Delta(X_i)_-$ is a empirical (unbiased) version of the measure $(\chi^+ - \chi^-)_-$, so $(\hat{\eta}_j, \hat{d}_j)_{j \in \mathcal{I}^-}$ with corresponding weights $(\hat{d}_i)_{i \in \mathcal{I}^-}$ is an estimate for the measure $\mu^-_{\eta,\Delta}$.

E Experimental setup and implementation details

To ensure a rigorous, reproducible, and equitable comparison between our Optimal Transport framework and the FairReg baseline, we strictly to the evaluation protocol established by Taturyan et al. [38].

E.1 Details on the Unaware Baselines (FairReg and FRAPPÉ)

We contextualize the performance and computational profiles of FairReg [38] and FRAPPÉ [14].

FairReg: Grid discretization and dual optimization. FairReg operates as an unlabeled post-processing algorithm that does not map inputs to a deterministic continuous prediction. Instead, it relies on discretization, defining a uniform grid $\hat{\mathcal{Y}}_L$ over a bounded interval $[-B, B]$ consisting of

$2L + 1$ discrete points. The algorithm learns a randomized prediction function over this support, parameterized by a softmax function where the hyperparameter β controls the degree of entropic regularization. To enforce DP, they formulate a dual stochastic convex program from this entropically regularized objective. This dual problem is then minimized using a stochastic first-order oracle, typically implemented via Accelerated Stochastic Approximation (AC-SA2) to control the expected norm of the gradient mapping.

FairReg: Iterations (T). In their theoretical framework and codebase, the hyperparameter T defines the total number of stochastic gradient evaluations. Because the stochastic gradient of the dual objective is computed by sampling one independent unlabeled feature vector X per step, T represents individual iterations rather than full passes over the dataset. To guarantee a fair, scale-invariant convergence analysis across datasets of varying sizes (N), we parameterize the baseline's optimization in terms of training epochs. Specifically, we set $T = 60 \times N$ for all experiments. This ensures the FairReg algorithm performs 60 epochs over the unlabeled data.

FRAPPÉ: Additive post-hoc modules and SGD. FRAPPÉ is a post-processing framework that achieves fairness by adding a learned post-hoc module, denoted $T_{PP}(X)$, to the unfair scores output by a pre-trained base model. Because this additive transformation depends only on the covariates X, it does not require access to the sensitive attribute at inference time. In our experiments, matching the authors implementation, we model this post-hoc transformation using a 1-layer Multi-Layer Perceptron (MLP) with 64 hidden units. The module is trained using stochastic gradient descent (specifically the Adam optimizer with a learning rate of 0.01) to minimize a combination of prediction loss and a Macximum Mean Discrepency (MMD) fairness regularizer with a Gaussian kernel. Because FRAPPÉ relies on gradient-based optimization rather than closed-form mappings, setting the fairness penalty parameter (λ) to high values to enforce parity causes optimization instability, leading to the degraded predictive utility observed in our trade-off analyses.

E.2 General experimental details

Data preprocessing and splits. For all real-world datasets, the continuous target variable Y is scaled linearly to the interval $[-1, 1]$. For all experiments, we evaluate out-of-sample generalization by splitting the data into a training set and a hold-out test set (using an $80\% - 20\%$ split), previously stratified by the sensitive attribute S to preserve group prior probabilities.

Base models and estimators. To isolate the effect of the fairness interventions and ensure differences in performance are attributable to the post-processing algorithms, all methods share the same base estimators. Specifically, the unconstrained base regressor $\hat{\eta}(X)$ is instantiated as a standard Ordinary Least Squares (OLS) Linear Regression model (we provide additional base model ablations study in Section G.6). The probabilistic sensitive attribute classifier $\hat{\mathbb{P}}(S \mid X)$, which is required by both FairReg and our Unaware OT framework to estimate the local density ratios, is instantiated as a Logistic Regression model trained for up to 2000 iterations to ensure convergence.

Hyperparameters. For the FairReg and FRAPPÉ baselines, we follow their original implementations and optimization heuristics, with full configuration details deferred to Section E.1. For our OT framework, the discrete optimal transport plan π^* is solved exactly. To construct the final continuous, out-of-sample predictor $\hat{f}_{fair}$ from the discrete OT assignments, we train a Random Forest regressor (with 200 trees) to map the 2-dimensional feature space $(\hat{\eta}(x), \hat{\Delta}(x))$ to the optimal fair targets y^*. The parameter τ in 1 is set to the value 10^{-6}.

Statistical significance and error bars. To assess the stability and statistical significance of our results, all experiments are aggregated over multiple independent trials (10 runs unless otherwise specified). The source of random variability depends on the experimental setting: for the synthetic dataset (Section F.2), the random seed dictates the underlying data generation process; for real-world datasets (Sections 4 and G), the random seed controls the stratified train/validation/test splits. In all summary tables, the reported $\pm$ values represent the empirical standard deviation calculated across these independent runs (Sections G.3, G.6 and H). Similarly, in all visualizations (such as the accuracy-fairness trade-off and confounding sweep curves), the shaded regions represent ± 1 standard deviation.

F Experiments on synthetic datasets

We use two distinct synthetic datasets to empirically illustrate the exact geometric behavior of the optimal transport relaxations in both the aware and unaware settings: a non-linear confounded dataset designed to visualize the effect of the relaxations on the individual predictions, and a linearly confounded dataset designed to sweep the confounding strength and evaluate the accuracy-fairness trade-off across different levels of bias.

F.1 Empirical illustration of the geometric relaxations

Experimental setup and data generation. To isolate and visualize the geometric behavior of the optimal transport relaxations, we use a 1-dimensional synthetic dataset designed to introduce a non-linear confounding effect. We generate $N = 5000$ independent samples according to the following data generating process:

- **Sensitive attribute:** We sample the binary group uniformly at random $S \sim \text{Bernoulli}(0.5)$, and define the corresponding Rademacher variable $Z_S = 2(S - 0.5) \in \{-1, 1\}$.
- **Confounded feature:** We construct the continuous feature $X \in \mathbb{R}$ as a linear combination of the sensitive group and an independent Gaussian noise: $X = \gamma Z_S + \sqrt{1-\gamma^2}\mathcal{N}(0, 1)$, where the confounding strength is set to $\gamma = 0.6$.
- **Target variable:** The continuous outcome Y follows a non-linear S-curve dependence on X with additive Gaussian noise: $Y = 3.0 \tanh(1.5X) + \mathcal{N}(0, 0.5^2)$.

To remove finite-sample estimation artifacts and evaluate the pure geometry of the optimal transport framework, we bypass empirical risk minimization for the base models. Instead, we use oracles for the continuous conditional expectation $\eta(x) = 3.0 \tanh(1.5x)$ and the exact posterior probabilities $\mathbb{P}(S = 1 \mid X = x) = \sigma\left(\frac{2\gamma x}{1-\gamma^2}\right)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

Finally, to ensure an equitable comparison between the $\mathcal{W}_2$ and TV penalties across both the aware and unaware paradigms, we perform a numerical line search to identify the exact λ (or exact threshold τ for the aware TV case) that matches a fairness budget (e.g., preserving exactly 80%, 60%, or 20% of the initial unconstrained Wasserstein-2 unfairness).

Qualitative analysis of the unaware relaxations To complement the analysis of the marginal distributions provided in the main text, we visualize here the effect of the relaxations on the individual predictions.

As shown in Fig. 5, comparing the base ERM predictor against the optimal relaxed unaware predictors (matched to an 80% unfairness budget) highlights the geometric difference between the two penalties. The $\mathcal{W}_2$ penalty induces a global smooth interpolation, pushing all individuals closer to the fair center. Conversely, the TV penalty creates a flat plateau, resulting in an overlapping density spike in the center of the distribution while leaving the extreme predictions untouched.

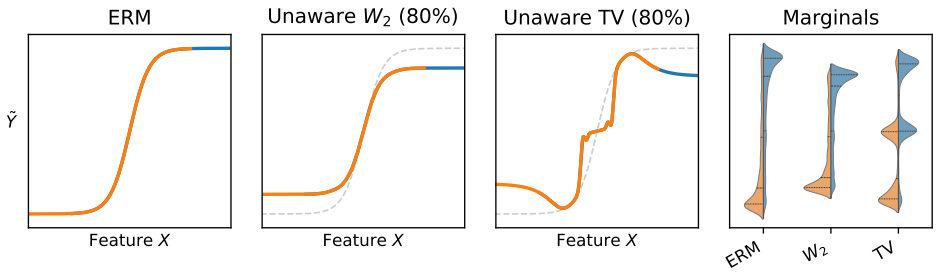

Figure 5: **Unaware relaxations.** Comparison of the Bayes predictor against the optimal relaxed unaware predictors under $\mathcal{W}_2$ and TV penalties, matched to an 80% unfairness budget.

Qualitative analysis of the aware relaxations To complement the unaware visualizations above, we empirically evaluate the exact same relaxations in the *aware setting* on our 1D synthetic dataset.

As shown in Fig. 6, the aware $\mathcal{W}_2$ penalty induces a smooth interpolation, shifting predictions homogeneously toward the central compromise. Conversely, the aware TV penalty acts as a hard-thresholding operator. This creates a highly discontinuous mapping: predictions merge if the pointwise disparity gap falls below the tolerated threshold, but keepthe unconstrained biased baseline if the gap is too large.

By varying the regularization parameter to transition the model from the unconstrained empirical risk minimizer (100% unfair) to strict Demographic Parity (0% unfair), we observe the evolution of the marginal distributions in Fig. 7. Consequently, the aware TV penalty fragments the intermediate marginal distributions into distinct, multi-modal subgroups.

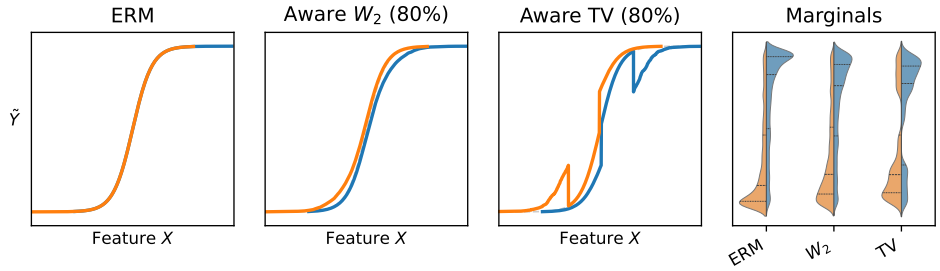

Figure 6: **Aware relaxations.** Comparison of the base ERM predictor against the optimal relaxed aware predictors under $\mathcal{W}_2$ and TV penalties. While $\mathcal{W}_2$ yields a smooth shift, the TV penalty applies a hard-thresholding, creating a discontinuous effect where predictions either perfectly merge or go back to the unfair ERM.

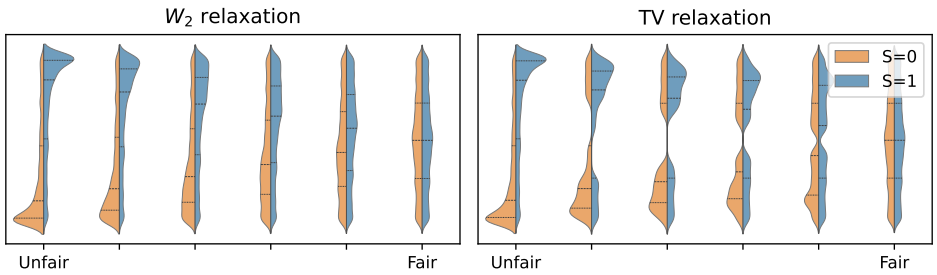

Figure 7: **Evolution of the aware marginal distributions.** Evolution of the predicted marginal distributions from Unfair to Fair. The discontinuous nature of the aware TV hard-thresholding creates fragmented intermediate distributions, contrasting with the smooth continuous shift of the $\mathcal{W}_2$ penalty.

F.2 Impact of confounding strenght and trade-offs

We generate n independent samples from a controlled 2-dimensional synthetic dataset.

Sensitive attribute and features. We sample the binary sensitive attribute uniformly at random $S \sim \text{Bernoulli}(0.5)$, and define the corresponding Rademacher variable $Z_S = 2(S - 0.5) \in \{-1, 1\}$. We construct a 2-dimensional continuous feature vector $X = (X_1, X_2)$. The first feature, X_1, acts as the confounded signal and is generated exactly as described in Section F.1. The second feature, X_2, acts as a "safe" predictive signal and is drawn from an independent Gaussian distribution

$$X_2 \sim \mathcal{N}(0, 1).$$

Target variable. The continuous outcome Y is generated as a simple linear function of both features, with an independent additive Gaussian noise ε

$$Y = 2X_1 - X_2 + \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, 0.5^2).$$

Role of the parameter γ. This specific design introduces a single source of unfairness governed by the parameter $\gamma \in [0, 1]$, which dictates the *confounding strength*. When $\gamma = 0$, X_1 is completely independent of S. As $\gamma \to 1$, X_1 becomes increasingly confounded with the demographic groups. Because Y is a function of X, the unconstrained ERM will always converge to the true coefficients $[2.0, -1.0]$, and its MSE will always equal the noise variance $\text{Var}(\varepsilon) = 0.25$, regardless of γ. This guarantees that any drop in accuracy observed during post-processing is exclusively attributable to the fairness constraints.

Accuracy-fairness trade-offs. We first fix $\gamma = 0.5$ to induce a moderate structural bias and evaluate the relaxed predictors. As shown in Fig. 8, the unconstrained ERM achieves the theoretical optimal MSE of 0.25 but exhibits high unfairness. As we tighten the fairness constraints, FAUST-U recovers the Pareto frontier of the optimal aware interpolations. Notably, the naive plug-in estimator (Plug-In W2) fails to achieve the same fairness levels, demonstrating the risk of relying on a hard proxy classification $\widehat{S}$ rather than the continuous group probabilities.

Impact of confounding strength. To evaluate robustness across different biases, we fix the constraint to exact DP (no relaxation) and vary the confounding strength γ from 10^{-2} to 10^{0} (Figs. 9 and 10). Note that because the empirical $\mathcal{W}_2$ distance is computed between finite samples there is a small positive value due to sampling noise even for fair models, which is why the models do not reach zero on the logarithmic scale. This reveals a characteristic of the unaware geometry: it requires an active demographic signal to construct the 2D pseudo-measures. When γ is low, S is independent of X, causing the estimated density ratio $\hat{\Delta}(x)$ to vanish. FAUST-U collapses the predictions toward the global mean (incurring a high MSE) to guarantee fairness.

However, as bias appears ($\gamma > 0.1$), the unaware geometry fully resolves. When there is high-confounding, FAUST-U suppresses the reliance on the biased feature while preserving the signal from X_2, converging to the optimal bounds of FAUST-A. Furthermore, the naive plug-in approach (Plug-In W2) is not robust to this confounding, even at moderate levels. Because it forces a hard classification of the sensitive attribute, it applies the optimal transport plan based on misaligned proxy labels. This results in a massive spike in true $\mathcal{W}_2$ unfairness, only finally converging to the Aware bounds when $\gamma \to 1$ and the proxy becomes an oracle. Conversely, the FairReg baseline shows instability, with MSE penalties early in the trajectory.

Variance analysis. This behavior at low confounding is confirmed by analyzing the within-group variance of the predictions (Fig. 11). At low γ, the variance of the FAUST-U predictions collapses significantly compared to the ERM, reflecting the prediction of the constant prediction in the absence of a reliable demographic signal. As the confounding signal strengthens ($\gamma \to 1$), all fair models reduce their variance to suppress X_1, but FAUST-U demonstrates the most stable convergence toward the aware optimum.

G Experiments on real-world datasets

To complement the real-world evaluations in the main text, we provide the full accuracy-fairness trade-off analysis for the Adult (Census Income) and plot the full plot to see the whole FRAPPÉ baseline on the Communities & Crime and Law School datasets .

G.1 Adult dataset

Adult dataset detail. The Adult dataset comprises individual-level records extracted from the 1994 U.S. Census database. It contains 48,842 total observations characterized by 14 demographic and employment-related attributes, including education level, marital status, occupation, and hours worked per week. While the usual prediction task for this dataset is binary classification (predicting whether an individual earns more than \$50K per year), [38] convert this into a regression task by using a numeric proxy from the features (age) as the continuous target Y.

The dataset contains demographic disparities: the population is approximately 66.9% male and 33.1% female. The racial composition of the dataset is also unbalanced, comprising approximately 85.4% White, 9.6% Black, 3.2% Asian-Pac-Islander, 1.0% Amer-Indian-Eskimo, and 0.8% Other. For this experiment, the sensitive attribute considered is gender.

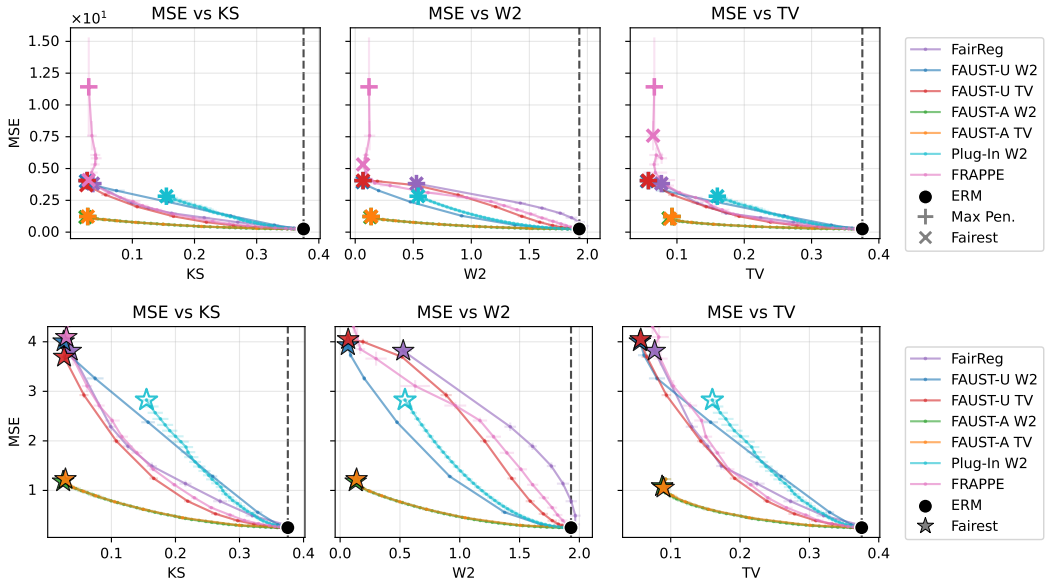

Figure 8: **Accuracy-fairness trade-offs on the synthetic dataset** ($\gamma = 0.5, n = 10,000$)**.** Relaxation trajectories from the unconstrained ERM (black dot) to the fair models. The top panel shows the full vertical scale capturing FRAPPÉ's extreme MSE degradation; the bottom panel provides a zoom on the remaining baselines. The unconstrained baseline converges to the noise limit (MSE $= 0.25$). Our exact mappings (FAUST-U $\mathcal{W}_2$ and FAUST-U TV) recover the optimal aware Pareto frontier (FAUST-A) and outperform FairReg and the noisy plug-in proxy (Plug-In W2).

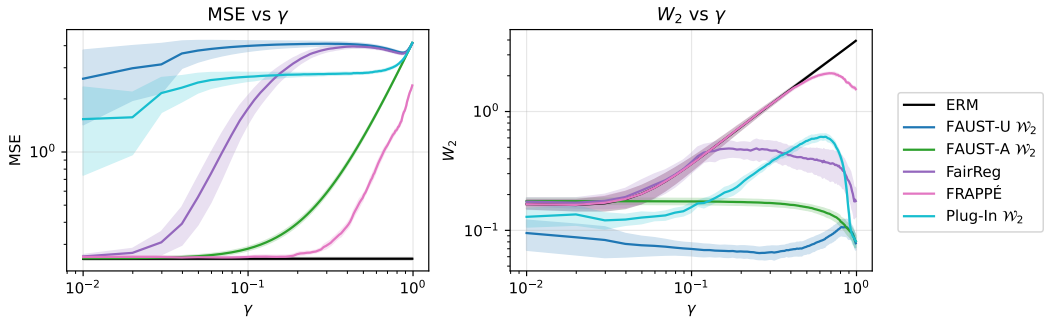

Figure 9: **Impact of confounding strength (Absolute).** Evolution of the MSE (left) and $\mathcal{W}_2$ unfairness (right) under exact DP. As bias increases ($\gamma \to 1$), FAUST-U converges to the theoretical aware optimum (FAUST-A), bounding the MSE degradation. In contrast, FairReg suffers from optimization instability, which degrades the MSE. FRAPPÉ shows instability, not minimize unfairness efficiently while showing a spike in MSE at high confounding levels.

Characteristics about the **Law School** and **Communities and Crime** datasets are provided in the main text Section 4.

G.2 Accuracy-Fairness trade-off

To complement our main evaluations, we present the full accuracy-fairness trade-off curves comparing our framework against the FRAPPÉ baseline [14]. As detailed in Section 4.1, FRAPPÉ acts as a relevant unaware post-processing baseline, but it relies on stochastic gradient descent to learn an additive post-hoc correction, rather than relying on exact continuous geometric mappings. We evaluated FRAPPÉ on all four datasets using the exact same base ERM regressors as our models, sweeping its penalty hyperparameter to trace its complete Pareto frontier.

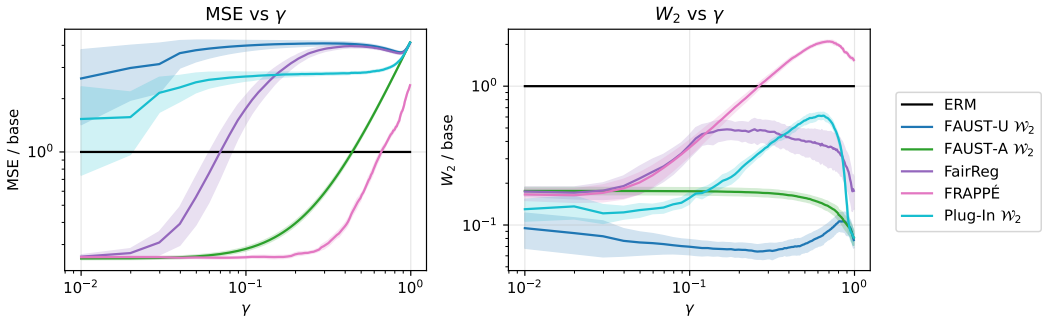

Figure 10: **Impact of confounding strength (Relative).** Evolution of the MSE and $\mathcal{W}_2$ unfairness expressed relative to the unconstrained ERM baseline. It highlight that both FairReg and FRAPPÉ's MSE drop even at moderate confounding levels ($\gamma \approx 0.1$), whereas FAUST-U maintains stability and accurately tracks the theoretical aware bound across all bias intensities.

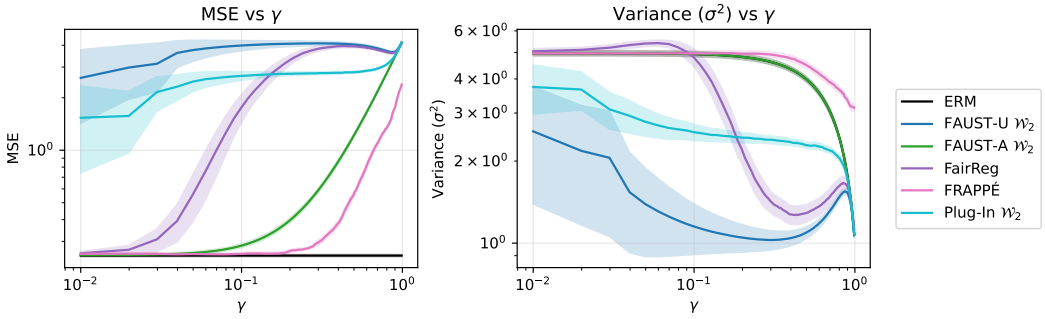

Figure 11: **Within-Group Variance Analysis.** Evolution of the MSE (left) and within-group prediction variance (right) across varying confounding strengths (γ). At low γ, FAUST-U collapses the prediction variance to act as a constant predictor, reflecting the uninformative demoigrphic signal $\hat{\Delta}(x)$. As bias increases, FAUST-U resolves the 2D geometry and recovers the variance of the optimal FAUST-A model, whereas baselines like FRAPPÉ show an artificizl variance into the predictions to satisfy gradient penalties.

Analysis of the trade-off geometry. As illustrated by the Pareto frontiers in Fig. 12, the generated curves show that FRAPPÉ suffers from optimization instability. As the fairness constraint tightens, FRAPPÉ's frontier forms a "hook", its MSE degrades without enforcing parity. This empirically validates the advantage of using closed-form mappings over gradient-based neural network approximations to enforce fairness constraints.

G.3 Evaluation of the fully constrained models

Experimental setup. Tables 2 to 4 present the performance of the fully constrained models relative to the unconstrained ERM (MSE > 1.00 indicates degradation; fairness metrics near 0.00 indicate perfect parity). Alongside FairReg, FRAPPÉ and FAUST-U, we evaluate two naive plug-in baselines. If S is unavailable at inference, an intuitive solution is to substitute it with an estimator $\widehat{S}$ (or $\widehat{p}(s \mid x)$) inside the aware optimal transport maps T_s, yielding

$$\widehat{Y}_{\text{hard}}(x) = T_{\widehat{S}(x)}(\eta(x)), \tag{17}$$

$$\widehat{Y}_{\text{soft}}(x) = \sum_s \widehat{p}(s \mid x)\, T_s(\eta(x)). \tag{18}$$

Results. On the Adult dataset (Table 2), FAUST-U consistently matches or outperforms FairReg across all metrics. While FairReg achieves the lowest fairness measure in terms of KS-distance between the predictions of the group, at equality with our FAUST-U method, it still suffers from a

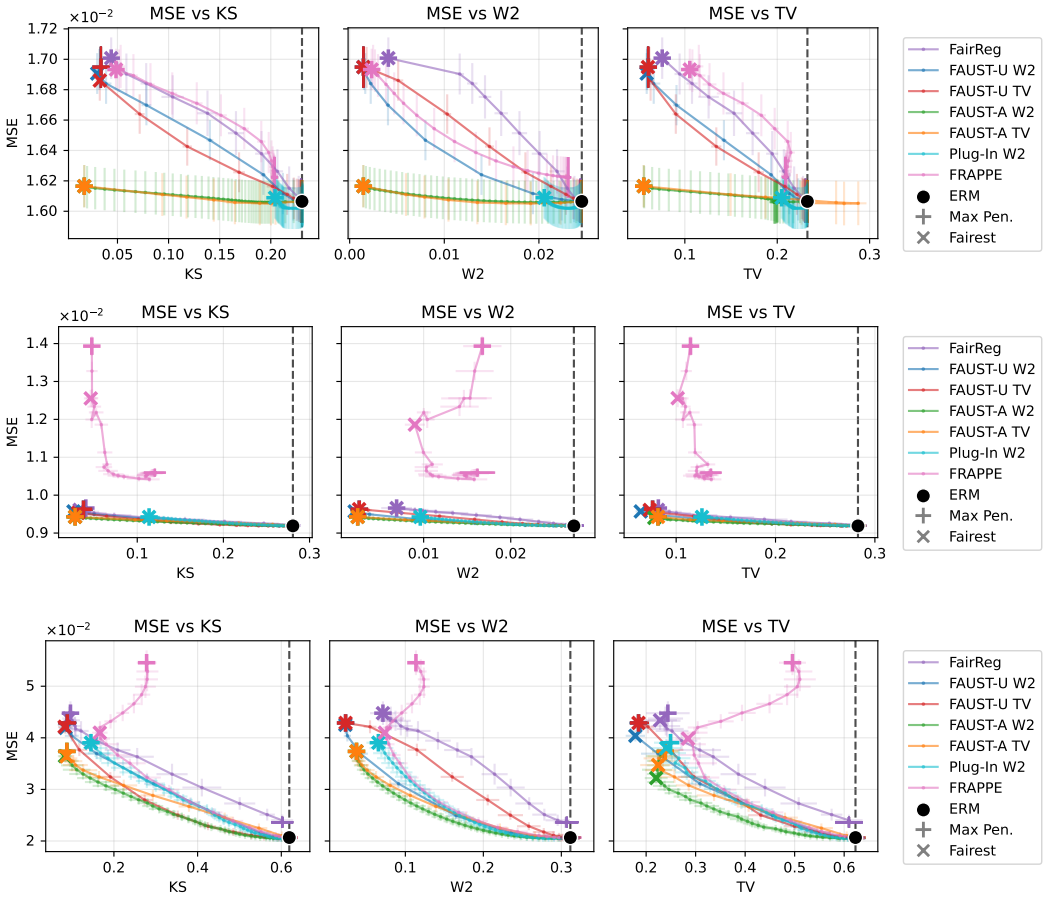

Figure 12: **Extended Accuracy-Fairness Trade-offs.** Relaxation trajectories across the evaluated datasets (from top to bottom: Adult, Law School and Communities and Crime). The light pink curve represents the FRAPPÉ baseline. Across all datasets, FRAPPÉ struggles to reach the theoretical fairness optimum (the star markers) without severe predictive degradation.

high $\mathcal{W}_2^2$-unfairness (23% relative to the baseline), whereas FAUST-U reduces it to 7% for the exact same relative MSE cost (1.06). FAUST-U also matches FairReg on TV and KS metrics, proving that exact optimal transport naturally guarantees statistical parity without the discretization.

Failure of naive plug-in estimators. The naive plug-in approaches completely fail to enforce fairness. On the Adult dataset (Table 2), the $\mathcal{W}_2^2$ and TV metrics for $\widehat{Y}_{\text{hard}}$ and $\widehat{Y}_{\text{soft}}$ remain between 0.85 and 1.10, making them virtually indistinguishable from the biased ERM. Because aware optimal transport relies on exact group rankings, classification errors from $\widehat{S}$ prevent achieving fairness. In contrast, FAUST-U successfully enforces parity by mathematically absorbing this classification uncertainty directly into its global density weights, succeeding where plug-in estimators fails.

G.4 Robustness to estimation error

Because our framework is a post-processing algorithm operating in the unaware setting, its performance relies on the base estimators for the conditional expectation $\hat{\eta}(x)$ and the posterior group probabilities $\hat{\mathbb{P}}(S \mid X)$ used to construct $\hat{\Delta}(x)$. In practice, estimation errors in these quantities are coupled and stem from limited training data.

Finite-sample study. To quantify the sensitivity to these finite-sample estimation errors, we restricted the training sample size N used to fit the base estimators and the optimal transport map. We

Method	MSE ↓	W_2 ↓	TV ↓	KS ↓	$\text{KS}^{\mathcal{T}}$ ↓
ERM Unfair	1.00 ± 0.02	1.00 ± 0.04	1.00 ± 0.04	1.00 ± 0.03	1.00 ± 0.04
FairReg	1.06 ± 0.02	0.23 ± 0.07	0.26 ± 0.05	$\mathbf{0.17 \pm 0.04}$	0.16 ± 0.05
FRAPPÉ	1.06 ± 0.01	0.10 ± 0.02	0.34 ± 0.04	0.25 ± 0.01	0.21 ± 0.03
FAUST-U	1.06 ± 0.01	$\mathbf{0.07 \pm 0.02}$	$\mathbf{0.23 \pm 0.04}$	0.17 ± 0.04	$\mathbf{0.15 \pm 0.05}$
Plug-In hard	$\mathbf{1.00 \pm 0.02}$	0.85 ± 0.03	0.88 ± 0.03	0.88 ± 0.03	0.88 ± 0.03
Plug-In soft	1.00 ± 0.02	1.10 ± 0.04	0.90 ± 0.04	0.88 ± 0.04	0.88 ± 0.04
FAUST-A	1.01 ± 0.02	0.06 ± 0.03	0.13 ± 0.03	0.10 ± 0.02	0.08 ± 0.03

Table 2: **Relative performance on the Adult dataset.** Performance of the fully constrained models evaluated relative to the unconstrained ERM baseline. Values are reported as mean $\pm$ standard deviation across 10 independent random splits. For fairness metrics (W_2, TV, KS, $\text{KS}^{\mathcal{T}}$), lower values indicate better parity; for MSE, values > 1.00 indicate accuracy degradation. The best-performing unaware method is highlighted in **bold**, while statistically indistinguishable unaware methods ($p > 0.05$ via Wilcoxon signed-rank test) are in blue. FAUST-U significantly improves W_2 over FairReg without incurring additional MSE degradation. While the FRAPPE baseline achieves a competitive MSE and W_2 gap on this dataset, it struggles to minimize the discrete KS and TV metrics compared to the exact FAUST-U mapping.

Method	MSE ↓	W_2 ↓	TV ↓	KS ↓	$\text{KS}^{\mathcal{T}}$ ↓
ERM Unfair	1.00 ± 0.02	1.00 ± 0.04	1.00 ± 0.03	1.00 ± 0.03	1.00 ± 0.03
FairReg	1.06 ± 0.02	0.41 ± 0.18	0.16 ± 0.07	$\mathbf{0.13 \pm 0.06}$	0.11 ± 0.08
FRAPPÉ	1.52 ± 0.04	0.68 ± 0.11	0.30 ± 0.04	0.17 ± 0.03	0.15 ± 0.03
FAUST-U	1.05 ± 0.02	$\mathbf{0.11 \pm 0.03}$	$\mathbf{0.14 \pm 0.06}$	0.14 ± 0.07	$\mathbf{0.10 \pm 0.07}$
Plug-In hard	1.03 ± 0.02	0.37 ± 0.05	0.43 ± 0.06	0.44 ± 0.06	0.42 ± 0.06
Plug-In soft	$\mathbf{1.02 \pm 0.02}$	0.46 ± 0.03	0.55 ± 0.07	0.56 ± 0.04	0.55 ± 0.07
FAUST-A	1.03 ± 0.02	0.11 ± 0.02	0.15 ± 0.05	0.13 ± 0.03	0.09 ± 0.03

Table 3: **Relative performance on the Law School dataset.** Metrics are normalized relative to the unconstrained ERM baseline, averaged across 10 random splits. Formatting follows the same conventions as the Adult dataset. FAUST-U systematically achieves superior geometric alignment (W_2) while matching the statistical parity (KS) of FairReg. In contrast, the naive plug-in estimators (Plug-In) fail to enforce meaningful fairness, demonstrating the fragility of relying on hard proxy labels. Additionally, the FRAPPE baseline exhibits instability under strict constraints, degrading predictive utility (relative MSE=1.52) while failing to enforce fairness ($W_2 = 0.68$).

Method	MSE ↓	W_2 ↓	TV ↓	KS ↓	$\text{KS}^{\mathcal{T}}$ ↓
ERM Unfair	1.00 ± 0.12	1.00 ± 0.07	1.00 ± 0.08	1.00 ± 0.09	1.00 ± 0.09
FairReg	2.00 ± 0.11	0.17 ± 0.08	0.37 ± 0.06	$\mathbf{0.16 \pm 0.04}$	$\mathbf{0.16 \pm 0.04}$
FRAPPÉ	2.37 ± 0.12	0.35 ± 0.03	0.81 ± 0.04	0.49 ± 0.07	0.47 ± 0.06
FAUST-U	1.90 ± 0.11	$\mathbf{0.09 \pm 0.04}$	$\mathbf{0.34 \pm 0.07}$	0.18 ± 0.07	0.16 ± 0.07
Plug-In hard	1.71 ± 0.12	0.23 ± 0.10	0.44 ± 0.05	0.27 ± 0.08	0.25 ± 0.09
Plug-In soft	$\mathbf{1.49 \pm 0.10}$	0.27 ± 0.10	0.42 ± 0.06	0.36 ± 0.07	0.34 ± 0.08
FAUST-A	1.64 ± 0.10	0.13 ± 0.04	0.38 ± 0.05	0.17 ± 0.04	0.15 ± 0.04

Table 4: **Relative performance on the Communities and Crime dataset**. Performance normalized relative to the ERM baseline across 10 random splits. FAUST-U systematically dominates the FairReg baseline, achieving both a lower relative MSE (1.92 vs 1.99) and significantly superior geometric alignment (W2=0.09 vs 0.20). This demonstrates that FairReg struggles to enforce parity on highly confounded datasets without severely degrading predictive utility. Finally, the FRAPPE baseline incurres the most MSE degradation (2.37) while failing to reach the fairness of FAUST-U or FairReg.

evaluated the evolution of the test MSE, the strict W_2 unfairness, and the tradeoff frontier across training sizes ranging from $N = 100$ up to the full dataset limits. All points report the mean and standard deviation across 10 independent random data splits.

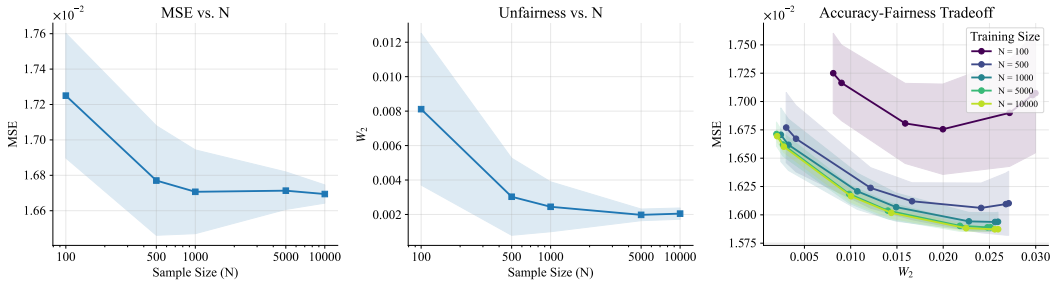

Figure 13: **Finite-sample performance on the Adult dataset.** *Left & Middle:* Test MSE and W_2 unfairness gap of the strictly fair FAUST-U model as a function of N. *Right:* Accuracy-fairness tradeoff Pareto frontiers. The Adult dataset exhibits stable corrections, the Pareto frontier going toward the optimal lower-left region as data increases.

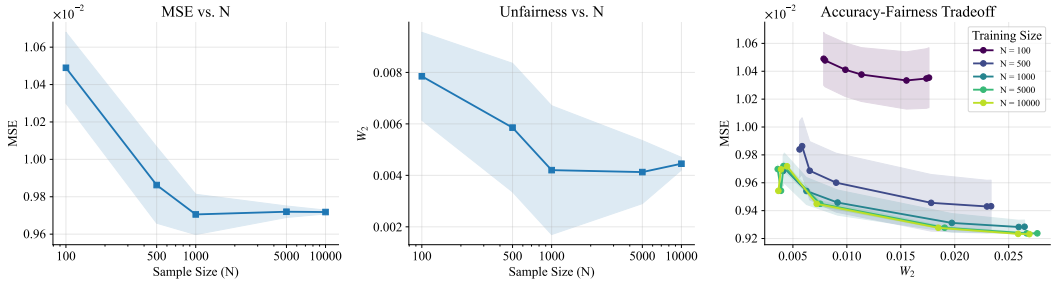

Figure 14: **Finite-sample performance on the Law School dataset.** Noticeably, under low-data regimes and strict fairness constraints (the rightmost points of the Pareto frontier), the lack of accurate density estimations causes a slight geometric distortion (a "hook" where MSE degrades without improving W_2). As N increases to 10,000 (light green curve), this artifact completely vanishes, yielding a strictly monotonic and optimal frontier.

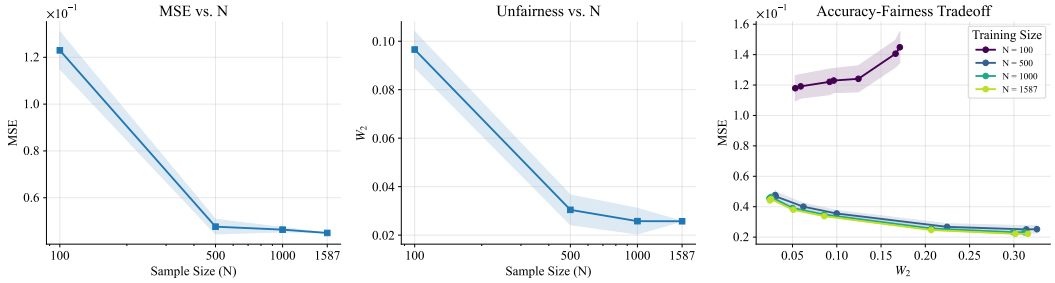

Figure 15: **Finite-sample performance on the Communities and Crime dataset.** Because this dataset contains 128 features, the $N = 100$ regime is nearly underdetermined, resulting in high estimation variance and a severely degraded frontier. However, with just $N = 500$ samples, the base estimators stabilize, and the optimal transport maps successfully enforce parity, rapidly converging to the optimal frontier up to the full dataset size ($N = 1587$).

Analysis of the results. Across all datasets, the results demonstrate that our OT framework is highly robust to estimation errors. Even with low data (e.g., $N = 100$) where the base estimators $\hat{\eta}$ and $\hat{\Delta}$ are noisy, the discrete optimal transport map enforces corrections, reducing the W_2 unfairness compared to the unconstrained ERM without impactong utility. While extreme noise can induce loss in performance under exact fairness constraints, such in the tail of the Law School Pareto frontier (Fig. 14),the overall tradeoffs remain resilient. Furthermore, the framework exhibits sample efficiency:

it requires very little data to overcome this noise and recover the optimal population-level mapping. For instance, despite the Communities dataset containing 128 features (Fig. 15), $N = 500$ samples are sufficient for the 2D pseudo-measures to resolve and match the full-dataset Pareto frontier.

Noise injection study. To further isolate the sensitivity of the post-processing framework to estimation errors, we conducted an artificial noise injection study. We introduced increasing levels of independent Gaussian noise ($\sigma \in [0.0, 0.5]$) into the base estimators $\hat{\mathbb{P}}(S \mid X)$ and $\hat{\eta}(X)$ separately, while keeping the fairness constraint strict ($\lambda \to \infty$). As shown in Fig. 16, the framework demonstrates strong robustness to miscalibration.

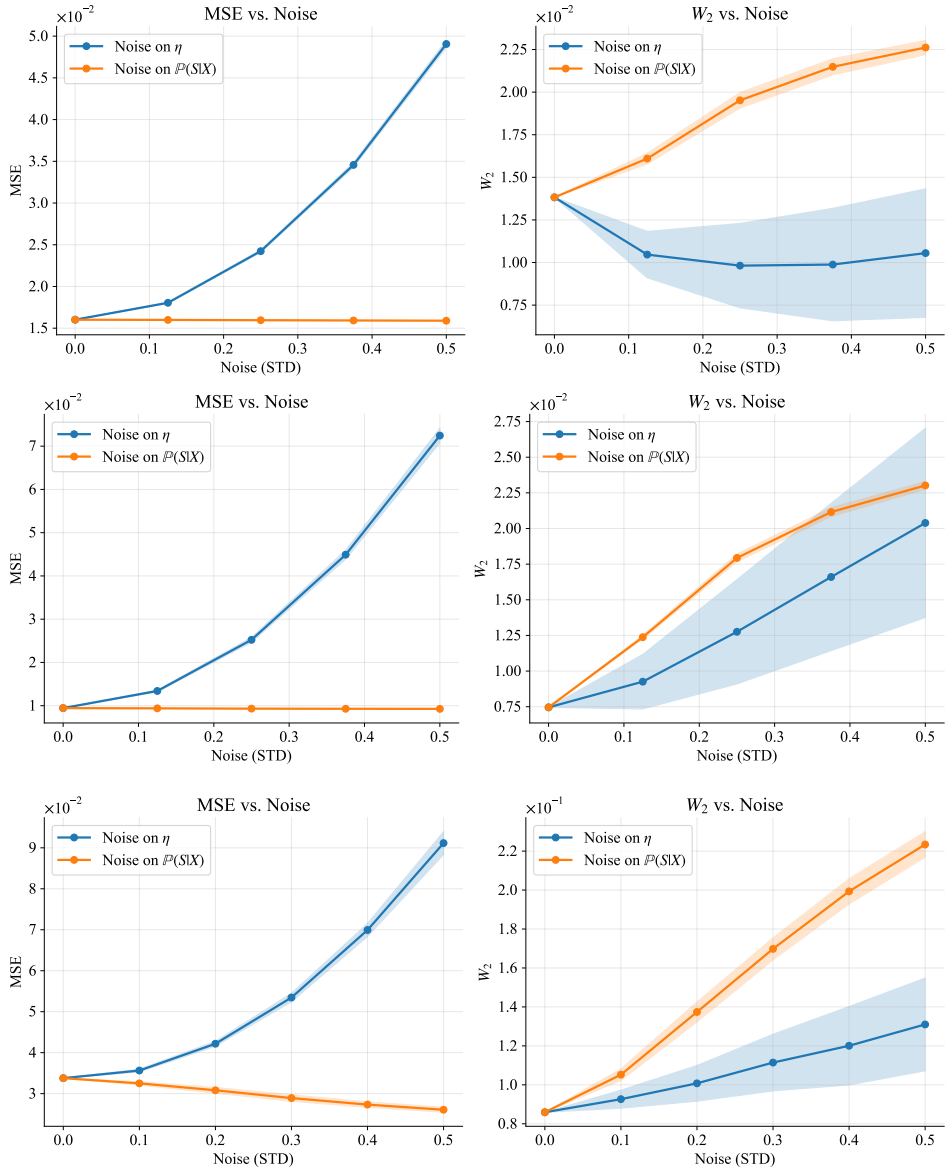

Figure 16: **Isolated metric degradation under estimator noise.** Evolution of the MSE (left column) and $\mathcal{W}_2$ gap (right column) of the fair FAUST-U model when artificial Gaussian noise is injected into either $\hat{\eta}$ (blue) or $\hat{\mathbb{P}}(S \mid X)$ (orange). The rows correspond to the Adult, Law School and, Communities and Crime datasets, respectively. Shaded regions represent ± 1 standard deviation across 5 independent runs.

When noise is applied to the posterior $\hat{\mathbb{P}}(S \mid X)$ (orange curves), the vertical density of the 2D space (η, Δ) is artificially smoothed. Consequently, the empirical pseudo-distributions appear more aligned prior to transport, causing the algorithm to under-correct. Fairness smoothly degrades (the $\mathcal{W}_2$ gap increases toward the unconstrained baseline), but this error does not significantly alter the predictive utility, leaving the MSE stable.

Conversely, when noise affects the base predictions $\hat{\eta}$ (blue curves), the OT map aligns the noisy distributions, maintaining a stable $\mathcal{W}_2$ gap near zero. However, the MSE degrades because the optimal transport map propagates the noisy base signals to satisfy the constraint.

Crucially, in both scenarios and across all datasets, the degradation is contzined. The framework does not fail under significant perturbation (up to $\sigma = 0.5$), confirming its viability and stability for real-world applications where base estimators are imperfect.

G.5 Sensitivity of fairness metrics to discretization grids

To evaluate DP on continuous predictions, metrics such as the Kolmogorov-Smirnov (KS) distance and the discrete Total Variation (TV) require binning the predictions into a finite grid. To ensure our empirical conclusions are robust to this evaluation hyperparameter, we tested the sensitivity of the metrics across fixed linear grids of 20, 50, and 100 bins, as well as an adaptive 50-bin setup (quantile-based binning to ensure equal probability mass per bin).

As illustrated in Fig. 17, the absolute values of the metrics, particularly TV, are sensitive to the chosen discretization, increasing as the evaluation grid becomes finer. However, despite this absolute variance, the relative ranking of the methods is invariant across all datasets. Across every tested configuration, our FAUST-U framework dominates both the unconstrained ERM and the FRAPPÉ baseline. Furthermore, it consistently matches or outperforms the grid-based baseline (FairReg), which explicitly optimizes for a discretized grid, remaining close to the theoretical optimal bound established by the Aware OT model. This confirms that our performance gains are highly robust to the choice of the evaluation grid.

G.6 Robustness across different base estimators

To ensure the performance of the post-processing methods is not an artifact of the initial model class, we benchmark the fully constrained frameworks across three distinct unconstrained base estimators: Ordinary Least Squares (Linear), Random Forests (RF), and Gradient Boosting Machines (GBM).

Methodology. For each base estimator, we extract the unconstrained predictions (ERM) and process them using both the FairReg baseline and our FAUST-U framework. To ensure a rigorous comparison, we report the mean and standard deviation across 10 random train/test splits of size. For the three unaware methods, the strictly superior performance is highlighted in **bold**. To account for variance, we apply a Wilcoxon signed-rank test; if a method underperforms the best method but the difference is not statistically significant ($p > 0.05$), it is highlighted in blue.

Metrics. We recall that $\text{KS}^{\mathcal{T}}$ is a discretized approximation of the Kolmogorov-Smirnov (KS) distance, computed as the maximum absolute c.d.f. difference across a discrete 50-bin grid $\mathcal{T}$. We also report the continuous Wasserstein-2 ($\mathcal{W}_2$) and Total Variation (TV) distances. The Total Variatioin is computed over 50 bins.

Base model configurations. To ensure reproducibility, we use the following hyperparameters for our unconstrained base estimators (implemented via scikit-learn):

- **Linear**: Ordinary Least Squares with an intercept (and Logistic Regression with `max_iter=2000` for auxiliary classification).
- **Random Forest (RF)**: 100 trees restricted to a minimum of 20 samples per leaf (`min_samples_leaf=20`) to prevent severe overfitting.
- **Gradient Boosting Machine (GBM)**: 100 boosting stages with a maximum tree depth of 3 (`max_depth=3`).

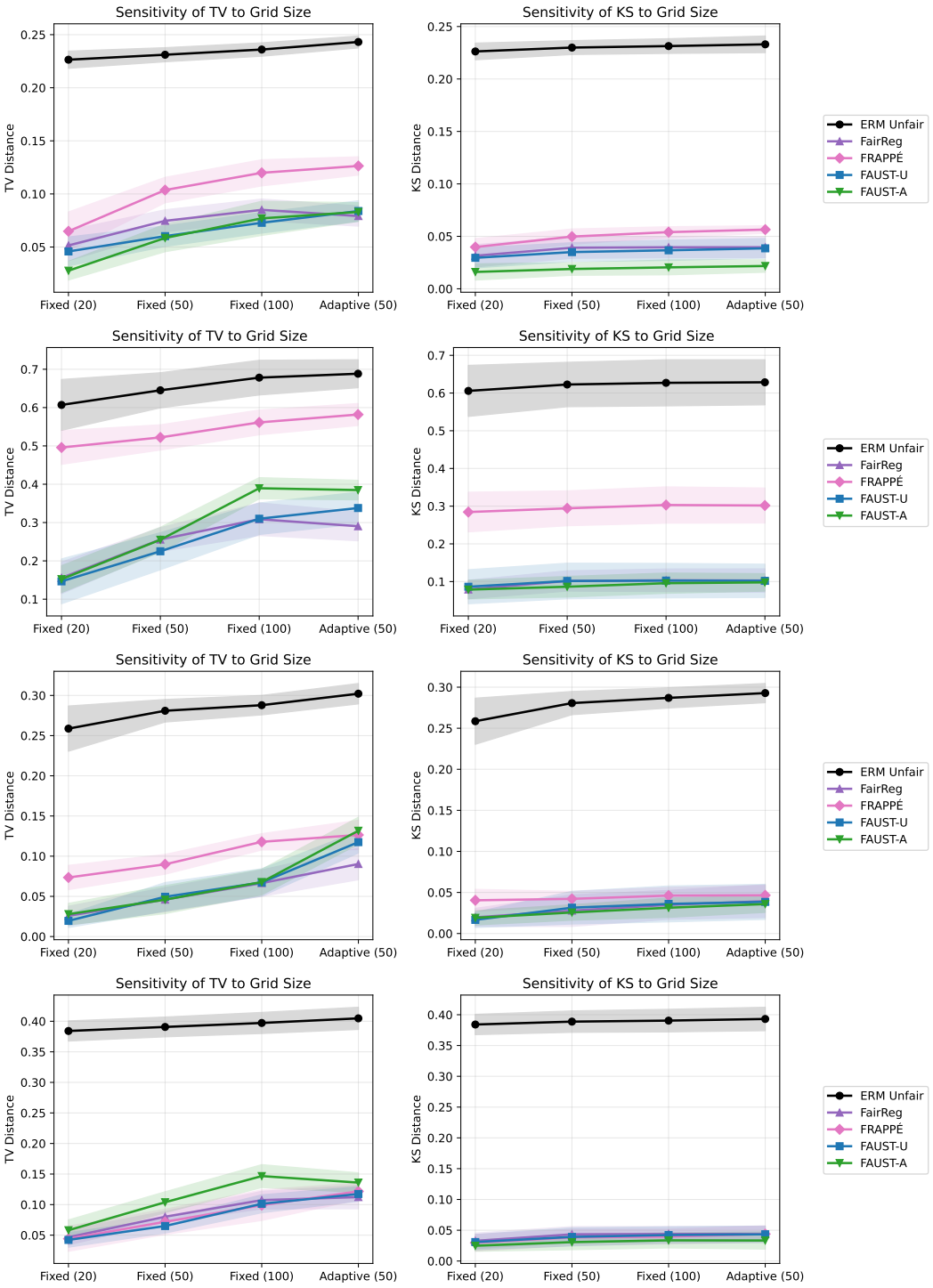

Figure 17: **Sensitivity of TV and KS metrics to evaluation bins.** While the absolute values of the metrics (especially the discrete TV) change with grid resolution, the relative performance of the methods remains stable across all configurations. The rows show the Adult, Communities and Crime, Law School, and Synthetic datasets, from top to bottom. Across all datasets, FAUST-U outperforms FRAPPÉ and matches or exceeds the FairReg baseline. Crucially, FAUST-U tracks the theoretical optimal bound (FAUST-A) regardless of the chosen evaluation grid.

Performance and interpretation. As shown in Tables 5 to 7, FAUST-U demonstrates robust generalization across all model classes. On the individual-level datasets (Adult and Law School), FAUST-U systematically achieves superior W_2 and $\text{KS}^{\mathcal{T}}$.

On the Communities dataset (Table 7), FAUST-U proves even more dominant. While it maintains highly competitive predictive utility (e.g., an MSE of 0.045 vs. 0.043 for FairReg on the Linear model), it successfully enforces the geometric fairness constraint. FairReg struggles to align the distributions, leaving a W_2 gap of 0.073 for the Linear model, whereas FAUST-U successfully reduces it to 0.031.

Theoretical validation of the MSE degradation. The main observation in the Adult and Law School results is that the empirical Mean Squared Error (MSE) appears unchanged between the Unfair ERM and the exact fair OT predictors (e.g., remaining at 0.01 or 0.02). This is not an anomaly, but a direct empirical validation of the theoretical bounds of Wasserstein fair regression.

The theoretical maximum MSE degradation required to achieve exact DP is bounded by the squared Wasserstein distance (W_2^2) between the initial biased distributions. For the Law School dataset, the initial W_2 gap is approximately 0.03. The exact geometric cost to repair this disparity is therefore bounded by $0.03^2 = 0.0009$. Because the targets are scaled to $[-1, 1]$, the base ERM risk is around 0.010. Adding the absolute maximum fairness penalty yields an expected fair risk of 0.0109. Consequently, the degradation is practically invisible at two decimal places. This proves that Optimal Transport successfully enforces parity with the absolute minimum possible degradation to accuracy.

Conversely, for the Communities and Crime dataset, the initial baseline disparities are severe ($W_2 \approx 0.31$). The theoretical geometric cost to achieve parity is therefore bounded by $0.31^2 \approx 0.096$. As expected, the empirical MSE degradation on this dataset is highly visible (jumping from 0.022 for the ERM to 0.045 for FAUST-U), tracking the quadratic cost bound.

Base Model	Method	MSE ↓	W_2 ↓	$\text{KS}^{\mathcal{T}}$ ↓
LINEAR	ERM Unfair	0.016 ± 0.000	0.024 ± 0.001	0.226 ± 0.010
	FairReg	0.017 ± 0.000	0.005 ± 0.002	0.037 ± 0.011
	FRAPPÉ	0.018 ± 0.000	0.002 ± 0.000	$\mathbf{0.022 \pm 0.008}$
	FAUST-U	$\mathbf{0.017 \pm 0.000}$	$\mathbf{0.002 \pm 0.000}$	0.034 ± 0.013
	FAUST-A	0.016 ± 0.000	0.002 ± 0.001	0.018 ± 0.008
RF	ERM Unfair	0.015 ± 0.000	0.020 ± 0.001	0.168 ± 0.011
	FairReg	$\mathbf{0.016 \pm 0.000}$	0.004 ± 0.001	0.021 ± 0.006
	FRAPPÉ	0.018 ± 0.000	0.005 ± 0.001	0.024 ± 0.011
	FAUST-U	0.016 ± 0.000	$\mathbf{0.003 \pm 0.000}$	$\mathbf{0.020 \pm 0.005}$
	FAUST-A	0.015 ± 0.000	0.003 ± 0.001	0.024 ± 0.011
GBM	ERM Unfair	0.015 ± 0.000	0.021 ± 0.001	0.198 ± 0.008
	FairReg	$\mathbf{0.016 \pm 0.000}$	0.006 ± 0.001	0.020 ± 0.009
	FRAPPÉ	0.017 ± 0.000	0.005 ± 0.001	0.025 ± 0.013
	FAUST-U	0.016 ± 0.000	$\mathbf{0.004 \pm 0.001}$	$\mathbf{0.019 \pm 0.006}$
	FAUST-A	0.015 ± 0.000	0.004 ± 0.001	0.019 ± 0.004

Table 5: **Robustness benchmark on the Adult dataset.** FAUST-U consistently drives geometric unfairness ($\mathcal{W}_2$) to near-zero levels across all base models (e.g., 0.002 for the Linear model),tracking the theoretical aware optimum (FAUST-A). While FairReg has a better MSE on the non-linear estimators (RF and GBM), FAUST-U maintains competitive predictive utility while having superior $\mathcal{W}_2$ alignment. Conversely, FRAPPÉ exhibits systematic MSE degradation across all configurations. Means and standard deviations are computed across 10 runs. The best performing unaware method is in **bold**; statistically indistinguishable results (Wilcoxon signed-rank test, $p > 0.05$) are highlighted in blue.

Base Model	Method	MSE ↓	W_2 ↓	$\text{KS}^{\mathcal{T}}$ ↓
LINEAR	ERM Unfair	0.009 ± 0.000	0.028 ± 0.001	0.278 ± 0.013
	FairReg	0.010 ± 0.000	0.007 ± 0.003	$\mathbf{0.032 \pm 0.022}$
	FRAPPÉ	0.025 ± 0.007	0.014 ± 0.006	0.034 ± 0.017
	FAUST-U	$\mathbf{0.010 \pm 0.000}$	$\mathbf{0.003 \pm 0.001}$	0.033 ± 0.021
	FAUST-A	0.010 ± 0.000	0.004 ± 0.001	0.028 ± 0.012
RF	ERM Unfair	0.008 ± 0.000	0.026 ± 0.002	0.209 ± 0.022
	FairReg	0.009 ± 0.000	0.008 ± 0.005	0.044 ± 0.011
	FRAPPÉ	0.017 ± 0.005	0.010 ± 0.003	$\mathbf{0.035 \pm 0.012}$
	FAUST-U	$\mathbf{0.009 \pm 0.000}$	$\mathbf{0.004 \pm 0.001}$	0.043 ± 0.013
	FAUST-A	0.008 ± 0.000	0.004 ± 0.001	0.033 ± 0.014
GBM	ERM Unfair	0.008 ± 0.000	0.026 ± 0.002	0.184 ± 0.021
	FairReg	0.008 ± 0.000	0.007 ± 0.004	$\mathbf{0.027 \pm 0.011}$
	FRAPPÉ	0.018 ± 0.002	0.011 ± 0.005	0.034 ± 0.028
	FAUST-U	$\mathbf{0.008 \pm 0.000}$	$\mathbf{0.003 \pm 0.001}$	0.028 ± 0.011
	FAUST-A	0.008 ± 0.000	0.004 ± 0.001	0.027 ± 0.005

Table 6: **Robustness benchmark on the Law School dataset.** FAUST-U consistently minimizes $\mathcal{W}_2$ unfairness across all base models while preserving predictive utility. Notably, FAUST-U matches the MSE of the FairReg baseline while tracking the theoretical aware optimum (FAUST-A). Conversely, FRAPPÉ exhibits severe MSE degradation across all configurations. Means and standard deviations are computed across 10 runs. The best performing unaware method is in **bold**; statistically indistinguishable results (Wilcoxon signed-rank test, $p > 0.05$) are highlighted in blue.

H Computational efficiency

We report here the average training time (in seconds) for each method across 10 independent runs on a standard CPU (Intel Core i7-9700K @ 3.60GHz). As shown in Table 8, the FAUST-U framework achieves massive computational speedups compared to the FairReg baseline, which relies on iterative stochastic gradient optimization. For instance, on the Communities dataset, FairReg takes on average 9.84 seconds to train, while FAUST-U takes only 0.35 seconds, yielding a speedup of over $28\times$.

Scalability with respect to sample size (N). To provide a precise empirical evaluation of the $O(N^3 \log N)$ worst-case complexity, we benchmarked the runtime of our exact discrete OT step as the number of unlabeled samples (N) increases (Table 9).

The runtime remains highly practical for dataset sizes up to $N = 20,000$, completing the entire post-processing pipeline in under 2 minutes. The cubic scaling becomes a computational bottleneck at $N = 50,000$, requiring approximately 45 minutes. Therefore, the exact method handles datasets up to 30,000 samples on standard hardware. Crucially, as a post-processing algorithm, this exact OT step only needs to be computed once on the available calibration dataset to construct the target mappings, rather than at every training epoch.

Scaling to massive datasets via Minibatch OT. For datasets significantly exceeding 30,000 points, resolving the exact OT plan becomes computationally prohibitive due to the $O(n^3 \log n)$ complexity. A classical and highly effective strategy to scale optimal transport is to use Minibatch OT [19], which estimates multiple barycenters and mappings independently on random subsets of the data, taking the final mapping as their average. Because the dimensionality of our mapped space is small (2D), minibatching provides excellent estimates. To empirically validate this, we conducted a study on a massive synthetic dataset of $N = 200,000$ samples, varying the minibatch size m from 50 to 10,000.

As shown in Table 10, the runtime scales exceptionally well, dropping from approximately 17 minutes ($m = 10,000$) to a fraction of a second ($m = 50$). More importantly, the approximation error is

Base Model	Method	MSE ↓	W_2 ↓	$\text{KS}^{\mathcal{T}}$ ↓
LINEAR	ERM	0.022 ± 0.003	0.310 ± 0.024	0.617 ± 0.065
	FairReg	$\mathbf{0.043 \pm 0.003}$	0.073 ± 0.031	0.112 ± 0.037
	FRAPPÉ	0.055 ± 0.006	0.058 ± 0.029	0.113 ± 0.051
	FAUST-U	0.045 ± 0.003	$\mathbf{0.031 \pm 0.013}$	$\mathbf{0.094 \pm 0.033}$
	FAUST-A	0.050 ± 0.006	0.062 ± 0.025	0.106 ± 0.031
RF	ERM	0.020 ± 0.001	0.281 ± 0.020	0.644 ± 0.054
	FairReg	$\mathbf{0.044 \pm 0.002}$	0.051 ± 0.028	0.131 ± 0.039
	FRAPPÉ	0.056 ± 0.008	0.061 ± 0.025	0.139 ± 0.063
	FAUST-U	0.045 ± 0.002	$\mathbf{0.024 \pm 0.010}$	$\mathbf{0.127 \pm 0.050}$
	FAUST-A	0.037 ± 0.002	0.027 ± 0.012	0.109 ± 0.062
GBM	ERM	0.021 ± 0.001	0.296 ± 0.023	0.630 ± 0.064
	FairReg	0.041 ± 0.003	0.051 ± 0.019	0.117 ± 0.033
	FRAPPÉ	0.056 ± 0.007	0.049 ± 0.018	0.114 ± 0.047
	FAUST-U	$\mathbf{0.041 \pm 0.002}$	$\mathbf{0.029 \pm 0.010}$	$\mathbf{0.109 \pm 0.029}$
	FAUST-A	0.037 ± 0.002	0.036 ± 0.013	0.112 ± 0.057

Table 7: **Robustness benchmark on the Communities and Crime dataset.** FAUST-U systematically achieves superior geometric alignment (W_2) and statistical parity ($\text{KS}^{\mathcal{T}}$) compared to the FairReg and FRAPPÉ baselines across all base model classes. For example, on the Linear base model, FairReg struggles to fully enforce parity (leaving a W_2 gap of 0.073), whereas FAUST-U strongly satisfies the geometric constraint ($W_2 = 0.031$) while maintaining highly competitive predictive utility (MSE of 0.045 vs. 0.043). FRAPPÉ consistently exhibits the most severe predictive degradation.

Dataset	FairReg	FRAPPÉ	FAUST-U	Plug-In	FAUST-A
Synthetic (10k)	48.77 ± 1.20	11.43 ± 0.40	7.23 ± 0.39	0.003 ± 0.001	0.001 ± 0.000
Adult	159.55 ± 1.59	0.27 ± 0.16	92.25 ± 4.46	0.031 ± 0.002	0.004 ± 0.000
Communities	9.84 ± 0.11	0.12 ± 0.15	0.35 ± 0.00	0.013 ± 0.001	0.007 ± 0.000
Lawschool	109.91 ± 0.79	0.23 ± 0.14	30.86 ± 0.85	0.050 ± 0.010	0.003 ± 0.000

Table 8: **Computational efficiency benchmark.** Average model training time (in seconds $\pm$ standard deviation) evaluated across 10 independent runs. By relying on exact, closed-form geometric reductions, the FAUST-U framework bypasses the iterative stochastic gradient optimization required by FairReg, yielding massive computational speedups (e.g., up to $26\times$ faster on the Communities dataset).

negligible: even with $m = 500$, **FAUST-U** enforces exact parity ($\mathcal{W}_2 \approx 0.006$) without suffering meaningful degradation in MSE, proving that the algorithm can scale to millions of samples.

N	**Exact OT Step (sec)** $\downarrow$	**Total Time (sec)** $\downarrow$
1,000	0.03± 0.00	0.21± 0.00
2,000	0.14± 0.05	0.47± 0.05
5,000	1.49± 0.60	2.36± 0.60
10,000	12.85± 0.16	14.76± 0.17
20,000	95.92± 1.00	100.21± 1.05
50,000	2705.94± 146.14	2718.59± 146.24

Table 9: **Empirical scalability of the exact discrete OT mapping.** Runtimes evaluated across increasing dataset sizes.

Minibatch Size (m)	**Num Batches** (K)	**OT Time (sec)** $\downarrow$	**MSE** $\downarrow$	$\mathcal{W}_2 \downarrow$
50	1,995	0.36	4.153	0.005
100	997	0.59	4.142	0.006
500	199	4.37	4.131	0.006
1,000	99	15.26	4.129	0.007
2,000	49	45.13	4.128	0.005
5,000	19	244.59	4.128	0.005
10,000	9	1021.26	4.127	0.005

Table 10: **Minibatch Optimal Transport on** $N = 200,000$**.** Averaging the continuous maps across minibatches reduces the computation time to a fraction of a second while maintaining strict parity.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the absyrtact and introduction we clearly state the paper's scope and contributions including a unifird framework for relaxed fair regression, closed-form solutions for $\mathcal{W}_2^2$ and TV penalties and empirical validations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We added a dedicated "Conclusion and Limitations" section discussing the restriction to binary sensitive attribute and the reliance on the calibration of the base estimators.
   '

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The key assumpotoons are clearly stated in the main text (e.g., Assumption 1 concerning the non-atomicity), and we did explicit reference to the appendix for proofs of the propositions and corollaries.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a detailed description of the algorithmic implementation in the main text and the experimental protocal is described more thouroughsly in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and instructions to reproduce th experiments are provided in the supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We explicit the experimental protocol, the baselines evaluated, dataset configurations and evaluation metrcs in the main text (Section 4). Further details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are reported with error bars (e.g., error bars across 10 runs with different random seeds) and the method for calculating the error bars is explained in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explicit the computational complexity of the framewrok in the main text (Section 3) and we provide an additional empirical time table in Table 8 in Section H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We strictly conform to the code of ethics. We addresse the "Bias and fairness" and "Discrimination" guidelines by introducing a mathematical framework explicitly designed to mitigate algorithmic bias and prevent models from exacerbating demographic disparities. In accordance with the "Impact Mitigation Measures," we disclose all essential elements for reproducibility, including full algorithmic details, open-source code, and the use of standard, publicly available datasets.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the ethical and societal philosophies underlying the choice of different fairness penalties in Section 2.4, adressing the broader impact of choosing specific techniques in algorithlmis fairness.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We introduce mathematical framework and post-processing algorithm for fair regression, it does not release any high-risk tool.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The real-world datasets used (Law School Admission, Communities and Crime, Adult) are standard public benchmarks dataset and properly cited. The open-source code for the baselines (FairReg) is also properly cited. Some of the baselines were implemented by us (e.g., [11]), but are base on the original papers that are cited in the main text and appendix.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduced all the details of our codes and synthetic datasets generated for the experiments in the main text and appendix, and we provide thourough instructions to reproduce the experiments in the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We did not use LLM as a component of the core methods or methodology of our research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.