WORST-CASE CONTROL OF FAIRNESS FOR CONTINUOUS SENSITIVE ATTRIBUTES

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair representation learning (FRL) seeks a representation that is informative for downstream prediction yet independent of a sensitive attribute. For a continuous sensitive attribute, existing methods measure the discrepancy between conditional and marginal distributions at each sensitive value and *average* it over the distribution of the sensitive attribute. Such an average can be small even when a fairness violation is confined to a low-probability range of sensitive values, and thus provides no uniform guarantee over the sensitive domain. We therefore propose the *supremum integral probability metric* (**supIPM**) to measure the worst-case discrepancy between the conditional and marginal distributions of the representation over sensitive values. supIPM controls the worst-case disparity of predictions, which we call L_∞-generalized demographic parity. We then propose **supFR**, an algorithm for learning representations that control worst-case fairness violations in downstream predictions. We establish a uniform convergence rate of the empirical supIPM and show that it yields a bound on downstream disparity for prediction heads of bounded Hölder norm. Experiments on tabular and graph benchmarks show that supFR attains prediction performance comparable to existing methods while substantially reducing the worst-case fairness violation.

1 INTRODUCTION

Representations learned without specifying a task at the training phase could be fruitfully used to construct prediction models in multiple downstream tasks (Bengio et al., 2013). To ensure that downstream prediction models are fair, Fair representation learning (FRL) seeks a representation that is independent of a given sensitive attribute, but yet informative for downstream prediction tasks (Zemel et al., 2013; Moyer et al., 2018; Creager et al., 2019; Song et al., 2019; Balunović et al., 2021; Gitiaux & Rangwala, 2021; Jovanović et al., 2023).

Most existing FRL methods have mainly focused on categorical sensitive attributes, where these methods try to align the distribution of the representation across each sensitive group. Adversarial learning is popularly employed in FRL (Edwards & Storkey, 2015; Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018). Alternatively, various integral probability metrics (IPMs) have been used to measure the degree of fairness violation of a given representation (Louizos et al., 2015; Kim et al., 2022; Deka & Sutherland, 2023). These methods, however, are not easily extended to continuous sensitive attributes such as age or income because the empirical distribution of a representation at each value of a continuous sensitive attribute is not easily estimated. We could discretize a continuous sensitive attribute, but the results are sensitive to the choice of bins (Kong et al., 2025), and data around the boundaries of each bin would not be fully utilized (Mary et al., 2019).

Algorithms to learn fair prediction models for continuous sensitive attributes without discretizing have been proposed by Mary et al. (2019); Grari et al. (2019; 2021); Jiang et al. (2022); Sohn et al. (2024), and an extension of these methods to FRL has been proposed by Kong et al. (2025). The shared idea of such methods is to control the average fairness violation of a given model or representation with respect to a continuous sensitive attribute. A critical limitation, however, in controlling the average fairness violation is that under-represented sensitive groups could be highly discriminated against (Kearns et al., 2018; Hashimoto et al., 2018).

The aim of this paper is to develop an FRL algorithm for continuous sensitive attributes which can control the worst-case fairness violation and thus avoid discrimination against under-represented

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

sensitive values. For this purpose, we propose a measure for the worst-case fairness violation of a given representation and develop an algorithm to learn a representation while controlling the worst-case fairness violation. Our proposed measure is the supremum of the IPM_s with respect to s, where IPM_s is the IPM between the distribution of a representation when the sensitive attribute S is equal to s and the marginal distribution, which we call ***supIPM***. For the choice of IPM, we consider ReLU-IPM of Park et al. (2026) since computation of supIPM can be done without much hamper and the worst-case fairness violations for a large class of prediction heads are well controlled. Theoretically, we show that a consistent estimator of supIPM can be obtained by combining ReLU-IPM and kernel density estimation. The worst-case fairness violation of any prediction model constructed on the fair representation is also asymptotically controlled as long as its prediction head is Hölder continuous.

Contributions.

- We propose supIPM as a measure of the worst-case fairness violation of a given representation for continuous sensitive attributes.
- We construct an estimator of supIPM by applying kernel density estimation techniques to ReLU-IPM, and prove that this estimator is consistent.
- We develop an algorithm to learn a fair representation while controlling the worst-case fairness violation, called supFR, and prove that the worst-case fairness violations of a large class of prediction models constructed on the fair representation learned by supFR are asymptotically controlled.
- We demonstrate empirically that supFR controls fairness violations across all of the domain of continuous sensitive attributes.

2 PRELIMINARIES

2.1 FAIR REPRESENTATION USING EIPM

We observe an input $\mathbf{X} \in \mathcal{X}$, an outcome $Y \in \mathcal{Y}$, and a continuous sensitive attribute S taking values in $\mathcal{S} \subset \mathbb{R}^r$. For notational simplicity, we only consider the case of $r = 1$ and let $\mathcal{S} = [0, 1]$. The proposed algorithm, however, can be easily modified for high-dimensional continuous sensitive attributes. See Remark 1 for further discussions.

An encoder $h : \mathcal{X} \to \mathcal{Z}$ maps $\mathbf{X}$ to a representation $\mathbf{Z}$ in the representation space $\mathcal{Z} \subset \mathbb{R}^d$. A prediction head $f : \mathcal{Z} \to \mathbb{R}$ maps the representation to a real-valued prediction, and $f \circ h$ denotes the composed prediction model which maps an input to a prediction value. We write $\mathbb{P}_S$ for the distribution of S, $\mathbb{P}_{\mathbf{Z}}$ for the distribution of $\mathbf{Z}$, and $\mathbb{P}_{\mathbf{Z}|S=s}$ for the conditional distribution of $\mathbf{Z}$ given $S = s$.

The discrepancy between $\mathbb{P}_{\mathbf{Z}|S=s}$ and $\mathbb{P}_{\mathbf{Z}}$ is measured using the integral probability metric (IPM) with a discriminator class $\mathcal{V}$, a class of real-valued discriminator functions on $\mathcal{Z}$. The IPM associated with $\mathcal{V}$ between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ (Müller, 1997; Sriperumbudur et al., 2012) is defined as

$$\text{IPM}_{\mathcal{V}}(\mathbb{P}, \mathbb{Q}) := \sup_{v \in \mathcal{V}} \left| \int v(\mathbf{z}) \, \mathrm{d}\mathbb{P}(\mathbf{z}) - \int v(\mathbf{z}) \, \mathrm{d}\mathbb{Q}(\mathbf{z}) \right|. \tag{1}$$

We say that $\text{IPM}_{\mathcal{V}}$ separates probability distributions if $\text{IPM}_{\mathcal{V}}(\mathbb{P}, \mathbb{Q}) = 0$ implies $\mathbb{P} = \mathbb{Q}$.

For a continuous sensitive attribute, Kong et al. (2025) define the expectation of integral probability metric (EIPM) as

$$\text{EIPM}_{\mathcal{V}}(\mathbf{Z}; S) := \int_{\mathcal{S}} \text{IPM}_{\mathcal{V}}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_{\mathbf{Z}}) \, \mathrm{d}\mathbb{P}_S(s) \tag{2}$$

and propose an FRL algorithm using EIPM called FREM. If $\text{IPM}_{\mathcal{V}}$ separates probability distributions, $\text{EIPM}_{\mathcal{V}}(\mathbf{Z}; S) = 0$ implies $\mathbb{P}_{\mathbf{Z}|S=s} = \mathbb{P}_{\mathbf{Z}}$ for $\mathbb{P}_S$-almost every s.

Given the training data $\mathcal{D} = \{(\mathbf{X}_i, S_i, Y_i)\}_{i=1}^n$ and $\mathbf{Z}_i = h(\mathbf{X}_i)$, FREM estimates the conditional distribution at each observed S_i by kernel weighting,

$$\widehat{\mathbb{P}}^{(-i),\gamma}_{\mathbf{Z}|S=S_i}(\cdot) := \sum_{j \neq i} \widehat{w}^{(-i)}_{\gamma}(j) \, \delta_{\mathbf{Z}_j}(\cdot), \qquad \widehat{w}^{(-i)}_{\gamma}(j) := \frac{K_\gamma(S_j, S_i)}{\sum_{r \neq i} K_\gamma(S_r, S_i)}, \tag{3}$$

where K_γ is a given kernel on $\mathcal{S}$ with the bandwidth $\gamma > 0$, $\delta_{\mathbf{z}}(\cdot)$ denotes the point mass at $\mathbf{z}$, and the superscript $(-i)$ indicates that observation i is left out. Comparing $\widehat{\mathbb{P}}^{(-i),\gamma}_{\mathbf{Z}|S=S_i}(\cdot)$ with the leave-one-out empirical marginal distribution $\widehat{\mathbb{P}}^{(-i)}_{\mathbf{Z}}(\cdot) = (n-1)^{-1}\sum_{j\neq i}\delta_{\mathbf{Z}_j}(\cdot)$ yields the estimator

$$\widehat{\mathrm{EIPM}}^{\gamma}_{\mathcal{V}}(\mathbf{Z}; S) := \frac{1}{n}\sum_{i=1}^{n} \mathrm{IPM}_{\mathcal{V}}\left(\widehat{\mathbb{P}}^{(-i),\gamma}_{\mathbf{Z}|S=S_i}, \widehat{\mathbb{P}}^{(-i)}_{\mathbf{Z}}\right). \tag{4}$$

With a general discriminator class, this estimator requires a separate maximization for calculating $\mathrm{IPM}_{\mathcal{V}}$ at every S_i, which is computationally demanding. FREM avoids this problem by taking $\mathcal{V}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS), for which each IPM is the MMD and has a closed-form kernel expression (Gretton et al., 2012; Kong et al., 2025). For a supervised loss ℓ and a penalty parameter $\lambda \geq 0$, FREM then trains the encoder h and the prediction head f by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(h(\mathbf{X}_i)), Y_i) + \lambda\, \widehat{\mathrm{EIPM}}^{\gamma}_{\mathcal{V}}(h(\mathbf{X}); S). \tag{5}$$

2.2 ReLU-IPM

It turns out that EIPM can control fairness violation of $f \circ h$ when f belongs to the discriminator class $\mathcal{V}$. When $\mathcal{V}$ is the RKHS with the Gaussian kernel, it contains only infinitely differentiable functions, so a prediction head of lower smoothness may not belong to $\mathcal{V}$. To overcome this restriction, Park et al. (2026) consider ReLU-IPM, an IPM whose discriminator class is indexed by a finite-dimensional parameter. ReLU-IPM controls the fairness violation of Hölder smooth prediction heads.

Let $\mathbb{B}^d := \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 \leq 1\}$ be the unit ball and let $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$ be the unit sphere. Park et al. (2026) define the ReLU discriminator class $\mathcal{V}_{\mathrm{R}}$ on $\mathbb{B}^d$ by

$$\mathcal{V}_{\mathrm{R}} := \left\{v_{\theta,\mu}(\mathbf{z}) = (\theta^\top \mathbf{z} + \mu)_+ : \theta \in \mathbb{S}^{d-1},\ \mu \in [-1, 1]\right\}, \tag{6}$$

where $(a)_+ := \max\{a, 0\}$. Despite its finite-dimensional parameterization of the discriminator class, ReLU-IPM separates probability distributions on $\mathbb{B}^d$, while keeping its computation tractable. Park et al. (2026) compute ReLU-IPM between two distributions by applying gradient ascent over (θ, μ), and projecting θ back onto the unit sphere in each update.

To specify the class of the prediction heads that ReLU-IPM controls, fix a smoothness index $\beta > 0$ and let k be the largest integer strictly smaller than β. Let $\partial^{\mathbf{a}}$ denote the partial derivative indexed by a multi-index $\mathbf{a} \in \mathbb{N}_0^d$ of order $|\mathbf{a}| = a_1 + \cdots + a_d$. Then the β-Hölder norm of a k times continuously differentiable function $v : \mathbb{B}^d \to \mathbb{R}$ and the β-Hölder ball on $\mathbb{B}^d$ are defined as

$$\|v\|_{\mathcal{H}^\beta} := \max\left\{\max_{|\mathbf{a}|\leq k}\sup_{\mathbf{z}\in\mathbb{B}^d}\left|\partial^{\mathbf{a}} v(\mathbf{z})\right|,\ \max_{|\mathbf{a}|=k}\sup_{\mathbf{z}\neq\mathbf{z}'\in\mathbb{B}^d}\frac{|\partial^{\mathbf{a}} v(\mathbf{z}) - \partial^{\mathbf{a}} v(\mathbf{z}')|}{\|\mathbf{z}-\mathbf{z}'\|_2^{\beta-k}}\right\} \tag{7}$$

and $\mathcal{H}^\beta_d := \left\{v \in C^k(\mathbb{B}^d) : \|v\|_{\mathcal{H}^\beta} \leq 1\right\}$. Park et al. (2026) prove that, for $d \geq 3$ and every $\beta > 0$ with $\beta \neq (d+3)/2$, there exists a constant C depending only on d and β such that, for all probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{B}^d$,

$$\mathrm{IPM}_{\mathcal{H}^\beta_d}(\mathbb{P}, \mathbb{Q}) \leq C \begin{cases} \mathrm{IPM}_{\mathcal{V}_{\mathrm{R}}}(\mathbb{P}, \mathbb{Q})^{\frac{2\beta}{d+3}}, & \beta < \frac{d+3}{2}, \\ \mathrm{IPM}_{\mathcal{V}_{\mathrm{R}}}(\mathbb{P}, \mathbb{Q}), & \beta > \frac{d+3}{2}. \end{cases} \tag{8}$$

This result implies that we can control Hölder-IPM by controlling ReLU-IPM, and thus control fairness violation of Hölder smooth prediction heads. However, computing $\widehat{\mathrm{EIPM}}^{\gamma}_{\mathcal{V}_{\mathrm{R}}}(\mathbf{Z}; S)$ requires adversarial learning for each S_i, making it computationally demanding to apply ReLU-IPM to FREM.

3 supIPM for fair representation learning

3.1 Uniform Fairness over the entire sensitive domain

Because the EIPM in (2) averages the IPM between $\mathbb{P}_{\mathbf{Z}|S=s}$ and $\mathbb{P}_{\mathbf{Z}}$ over $\mathbb{P}_S$, a large discrepancy of $\mathrm{IPM}_{\mathcal{V}}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_{\mathbf{Z}})$ at s in a region of small $\mathbb{P}_S$-probability could contribute little to $\mathrm{EIPM}_{\mathcal{V}}(\mathbf{Z}; S)$,

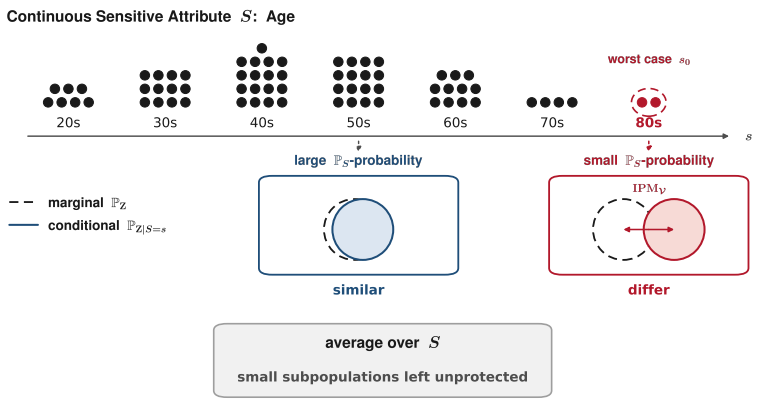


Figure 1: Overview of worst-case fairness violation.

and thus reducing $\mathrm{EIPM}_\mathcal{V}(\mathbf{Z}; S)$ may not reduce the $\mathrm{IPM}_\mathcal{V}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z})$ for all $s \in \mathcal{S}$. Figure 1 illustrates such a case: $\mathrm{IPM}_\mathcal{V}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z})$ could be large at s at which $\mathbb{P}_S$-probability is small even when $\mathrm{EIPM}_\mathcal{V}(\mathbf{Z}; S)$ is small, because a small EIPM need not imply a small discrepancy at every sensitive value. To control $\mathrm{IPM}_\mathcal{V}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z})$ for all s, in this paper we take the supremum instead of the expectation with respect to s. That is, we control $\sup_{s\in\mathcal{S}} \mathrm{IPM}_\mathcal{V}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z})$ instead of EIPM.

We define the supremum integral probability metric (supIPM) as

$$\begin{aligned}\mathrm{supIPM}_\mathcal{V}(\mathbf{Z}; S) &:= \sup_{s\in\mathcal{S}} \mathrm{IPM}_\mathcal{V}\left(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z}\right) \\ &= \sup_{s\in\mathcal{S}} \sup_{v\in\mathcal{V}} \left|\mathbb{E}[v(\mathbf{Z}) \mid S = s] - \mathbb{E}[v(\mathbf{Z})]\right|. \end{aligned} \tag{9}$$

By definition, the constraint $\mathrm{supIPM}_\mathcal{V}(\mathbf{Z}; S) \le \rho$ for $\rho \ge 0$ bounds $\mathrm{IPM}_\mathcal{V}(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_\mathbf{Z})$ by ρ at every $s \in \mathcal{S}$. Because the average never exceeds the supremum, $\mathrm{EIPM}_\mathcal{V}(\mathbf{Z}; S) \le \mathrm{supIPM}_\mathcal{V}(\mathbf{Z}; S)$, so a representation with a small supIPM also has a small EIPM.

For a bounded measurable function $g : \mathcal{X} \to \mathbb{R}$, Jiang et al. (2022) define generalized demographic parity (GDP) for a continuous sensitive attribute as

$$\mathrm{GDP}(g) := \mathbb{E}_S[|\mathbb{E}[g(\mathbf{X}) \mid S] - \mathbb{E}[g(\mathbf{X})]|]. \tag{10}$$

For $g = f \circ h$, $\mathrm{GDP}(g)$ averages $|\mathbb{E}[f(\mathbf{Z}) \mid S] - \mathbb{E}[f(\mathbf{Z})]|$ over S. Replacing the average by the supremum, as in (9), we define L_∞-generalized demographic parity (GDP_∞) as

$$\mathrm{GDP}_\infty(g) := \sup_{s\in\mathcal{S}} \left|\mathbb{E}[g(\mathbf{X}) \mid S = s] - \mathbb{E}[g(\mathbf{X})]\right|. \tag{11}$$

Note that $\mathrm{GDP}_\infty(f \circ h) \le \mathrm{supIPM}_\mathcal{V}(\mathbf{Z}; S)$ as long as $f \in \mathcal{V}$, and so we can control the fairness violation of $f(\mathbf{Z})$ for all $f \in \mathcal{V}$ in view of $\mathrm{GDP}(g)$ and $\mathrm{GDP}_\infty(g)$ by controlling $\mathrm{supIPM}_\mathcal{V}(\mathbf{Z}; S)$.

3.2 Choice of the Discriminator Class

It remains to choose the discriminator class $\mathcal{V}$. Note that $\mathrm{GDP}_\infty(f \circ h)$ for heads f with a bounded Hölder norm is bounded by $\mathrm{supIPM}_{\mathcal{H}_d^\beta}(\mathbf{Z}; S)$. However, computation of $\mathrm{supIPM}_{\mathcal{H}_d^\beta}(\mathbf{Z}; S)$ is not easy because a function in $\mathcal{H}_d^\beta$ is not easily parameterizable. To resolve this problem, we use ReLU-IPM, that is, we take $\mathcal{V} = \mathcal{V}_\mathrm{R}$ in (9) and control $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$. Since the discriminator class of ReLU-IPM has finite parameters, computation of $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$ can be done without much hamper. Moreover, the inequality (8) implies that $\mathrm{GDP}_\infty(f \circ h)$ for a Hölder smooth f is bounded by $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$ under regularity conditions.

ReLU-IPM has several advantages for supIPM compared to MMD. Computing supIPM with MMD requires $O(n^2)$ computation at each update. In contrast, $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$ only requires $O(n)$ computation at each update even though adversarial training is needed. Moreover, mini-batch learning is biased for supIPM with MMD (Lee et al., 2026), and $O(n^2)$ memory is required to store full kernel

matrix. In contrast, we show in Appendix B.2 that supIPM with the ReLU-IPM admits mini-batch learning.

Theoretically, the ReLU-IPM can control IPM with the discriminator class $\mathcal{H}_d^\beta$ which is much larger than that of MMD. Thus, controlling supIPM with $\mathrm{IPM}_{\mathcal{V}_\mathrm{R}}$ can guarantee the level of fairness violation for a large class of prediction heads. Empirical studies in Section 5 confirm this advantage.

3.3 Estimation of supIPM

Similar to EIPM, we can estimate $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$ by kernel techniques. Let

$$\widehat{w}_\gamma(i; s) := \frac{K_\gamma(S_i, s)}{\sum_{j=1}^n K_\gamma(S_j, s)}, \quad \widehat{\mathbb{P}}^\gamma_{\mathbf{Z}|S=s}(\cdot) := \sum_{i=1}^n \widehat{w}_\gamma(i; s)\, \delta_{\mathbf{Z}_i}(\cdot), \quad \widehat{\mathbb{P}}_\mathbf{Z}(\cdot) := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{Z}_i}(\cdot), \tag{12}$$

where K_γ is a given kernel. We propose to estimate $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S)$ by

$$\begin{aligned}\widehat{\mathrm{supIPM}}^\gamma_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S) &:= \sup_{s\in\mathcal{S}} \mathrm{IPM}_{\mathcal{V}_\mathrm{R}}\left(\widehat{\mathbb{P}}^\gamma_{\mathbf{Z}|S=s}, \widehat{\mathbb{P}}_\mathbf{Z}\right) \\ &= \sup_{s\in\mathcal{S}} \sup_{\theta\in\mathbb{S}^{d-1},\, \mu\in[-1,1]} \left|\sum_{i=1}^n \left(\widehat{w}_\gamma(i; s) - \frac{1}{n}\right)(\theta^\top \mathbf{Z}_i + \mu)_+\right|. \end{aligned} \tag{13}$$

We prove in Theorem 1 of Section 4 that this estimator converges to the population supIPM with a properly chosen kernel and bandwidth.

3.4 Fair Representation Learning with supIPM

To learn fair representations, we use the empirical supIPM of Section 3.3 as a fairness penalty. Let h_ϕ and f_ψ be the encoder and the prediction head with parameters ϕ and ψ, respectively. For $s \in \mathcal{S}$, $\theta \in \mathbb{S}^{d-1}$ and $\mu \in [-1, 1]$, let

$$\mathcal{L}(\phi, \psi; s, \theta, \mu) := \frac{1}{n}\sum_{i=1}^n \ell(f_\psi(h_\phi(\mathbf{X}_i)), Y_i) + \lambda \left|\sum_{i=1}^n \left(\widehat{w}_\gamma(i; s) - \frac{1}{n}\right)(\theta^\top h_\phi(\mathbf{X}_i) + \mu)_+\right|, \tag{14}$$

where ℓ is the supervised loss of Section 2.1 and $\lambda \geq 0$ controls the accuracy–fairness trade-off (Zhao & Gordon, 2022). We jointly train the encoder and the prediction head by solving $\min_{\phi,\psi} \mathcal{L}(\phi, \psi)$ with $\mathcal{L}(\phi, \psi) := \sup_{s,\theta,\mu} \mathcal{L}(\phi, \psi; s, \theta, \mu)$. Note that this problem is a Lagrangian formulation of the empirical supervised risk minimization subject to $\widehat{\mathrm{supIPM}}^\gamma_{\mathcal{V}_\mathrm{R}}(\mathbf{Z}; S) \leq \rho$.

To optimize this objective function, we alternate between the two steps. With (ϕ, ψ) fixed, we approximate the supremum over (s, θ, μ) in $\mathcal{L}(\phi, \psi)$ by compass search over s (Kolda et al., 2003) and projected gradient ascent over (θ, μ). With (s, θ, μ) fixed, we update (ϕ, ψ) by gradient descent. Because the compass search and the gradient ascent can converge to a local maximum, we use T initial values of s and M initial vectors of (θ, μ), respectively. We refer to this learning algorithm as supFR, which can be implemented efficiently using mini-batches (Appendix B.2). Further details and the algorithms are provided in Appendices B.1 and B.3, respectively.

Section 4 shows that, with high probability, a small empirical supIPM yields a small GDP_∞ on an interior region of $\mathcal{S}$ for every downstream head of bounded Hölder norm.

Remark 1 (Multivariate sensitive attributes)**.** *For $S \in \mathcal{S} \subset \mathbb{R}^r$ with $r > 1$, we use a multivariate kernel*

$$K_\gamma(s', s) = \gamma^{-r} \prod_{k=1}^r K((s'_k - s_k)/\gamma)$$

where $K(\cdot)$ is a univariate kernel in (12). We approximate the supremum in (13) by compass search over $\mathcal{S}$ along each coordinate, together with projected gradient ascent over (θ, μ). With T initial candidate vectors, each compass search iteration evaluates at most $T(2r+1)$ sensitive values (Kolda et al., 2003). Thus, supFR can be extended to multivariate sensitive attributes, with evaluations per iteration increasing linearly with r for fixed T.

4 THEORETICAL ANALYSIS

Theorem 1 establishes uniform convergence of the empirical supIPM to its population counterpart. Corollary 2 then bounds the GDP_∞ of prediction heads of bounded Hölder norm in terms of the empirical supIPM, as stated at the end of Section 3.4.

Theorem 1 and Corollary 2 are stated on the interior region $\mathcal{S}_n := [\epsilon_n, 1-\epsilon_n]$ of $\mathcal{S} = [0,1]$, where $0 < \epsilon_n < \frac{1}{2}$ and $\epsilon_n \to 0$. We restrict the suprema over s in (9), (11), and (13) to $\mathcal{S}_n$ and set $\gamma = \gamma_n$ for the empirical supIPM in (13). This gives

$$\begin{aligned} \mathrm{GDP}_{\infty,n}(g) &:= \sup_{s\in\mathcal{S}_n} \left|\mathbb{E}[g(\mathbf{X}) \mid S=s] - \mathbb{E}[g(\mathbf{X})]\right|, \\ \mathrm{supIPM}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) &:= \sup_{s\in\mathcal{S}_n} \mathrm{IPM}_{\mathcal{V}_\mathrm{R}}\big(\mathbb{P}_{h(\mathbf{X})|S=s}, \mathbb{P}_{h(\mathbf{X})}\big), \\ \widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) &:= \sup_{s\in\mathcal{S}_n} \mathrm{IPM}_{\mathcal{V}_\mathrm{R}}\big(\widehat{\mathbb{P}}^{\gamma_n}_{h(\mathbf{X})|S=s}, \widehat{\mathbb{P}}_{h(\mathbf{X})}\big). \end{aligned} \tag{15}$$

To bound the difference between $\widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S)$ and $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S)$ uniformly over a class $\mathcal{E}_n$ of encoders, we assume four conditions, stated in Appendix A.1. Assumption 1 requires $\mathbb{P}_S$ to have a density p_S with $0 < c_S \le p_S \le C_S < \infty$ on $\mathcal{S}$. Assumption 2 requires that the joint density of $(\mathbf{X}, S)$ is twice continuously differentiable in s, with integrable envelopes M_1 and M_2. Assumption 3 requires a bounded, symmetric and Lipschitz kernel K in $K_\gamma(s', s) = \gamma^{-1}K((s'-s)/\gamma)$, together with a bandwidth $\gamma_n \to 0$. Assumption 4 requires that the covering number of the encoder class $\mathcal{E}_n$ with respect to the sup norm is at most $(A_n/\eta)^{\mathfrak{v}_n}$ for every radius $\eta \in (0,1]$. To state the uniform convergence rate, let $0 < \delta < 1$ and define

$$\Gamma_{n,\delta} := (\mathfrak{v}_n + d)\log\frac{A_n n}{\gamma_n} + \log\frac{1}{\delta}, \qquad r_{n,\delta} := \gamma_n^2 + \sqrt{\frac{\Gamma_{n,\delta}}{n\gamma_n}}. \tag{16}$$

Theorem 1 (Uniform convergence rate of the empirical supIPM)**.** *Let Assumptions 1–4 hold. There exist constants $C, c > 0$, depending only on c_S, C_S, K, M_1 and M_2 such that for every $\delta \in (0,1)$ and every n with $\Gamma_{n,\delta} \le c\, n\gamma_n$, we have*

$$\sup_{h\in\mathcal{E}_n} \left|\widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) - \mathrm{supIPM}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S)\right| \;\le\; C\, r_{n,\delta}, \tag{17}$$

with probability at least $1-\delta$. If in addition A_n and $\mathfrak{v}_n$ are bounded, $\delta := \delta_n = n^{-c_1}$ for some $c_1 > 0$, $\gamma_n \asymp (\log n/n)^{1/5}$ and $\epsilon_n \asymp 1/\log n$, then $r_{n,\delta_n} = O\{(\log n/n)^{2/5}\}$.

Theorem 1 shows that the worst-case fairness violation on $\mathcal{S}_n$ can be transferred to the population one. That is, if $\widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) \le \rho$ for an encoder $h \in \mathcal{E}_n$, then $\mathrm{supIPM}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) \le \rho + Cr_{n,\delta}$ with probability at least $1-\delta$. Moreover, the bound is uniform over $\mathcal{E}_n$, so it applies to the encoder learned by supFR. In addition, since $\epsilon_n \to 0$, $\mathcal{S}_n$ converges to $[0,1]$, so every fixed sensitive value in the interior of $\mathcal{S}$ is eventually covered.

Theorem 1 and the inequality in (8) yield the following corollary. For $B > 0$ let $\mathcal{H}_d^\beta(B)$ be the Hölder ball of (7) with radius B, and let $C_{d,\beta}$ and $\alpha_{d,\beta}$ be the constant and exponent in (8), so that $\alpha_{d,\beta} = 2\beta/(d+3)$ for $\beta < (d+3)/2$ and $\alpha_{d,\beta} = 1$ for $\beta > (d+3)/2$.

Corollary 2. *Let Assumptions 1–4 hold. Let $d \ge 3$, and let C and c be the constants of Theorem 1. For every $\delta \in (0,1)$ and n with $\Gamma_{n,\delta} \le c\, n\gamma_n$, with probability at least $1-\delta$, simultaneously for all $h \in \mathcal{E}_n$, $B > 0$ and $\beta > 0$ with $\beta \ne (d+3)/2$,*

$$\sup_{f\in\mathcal{H}_d^\beta(B)} \mathrm{GDP}_{\infty,n}(f\circ h) \;\le\; C_{d,\beta}\, B \left\{\widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X});S) + Cr_{n,\delta}\right\}^{\alpha_{d,\beta}}. \tag{18}$$

This corollary implies that by using an encoder h with a small empirical supIPM, we can control the worst-case fairness violation over all prediction heads $f \in \mathcal{H}_d^\beta(B)$ constructed on h.

5 EXPERIMENTS

This section evaluates supFR on simulated data and on three tabular datasets. Experimental details are given in Appendices C and D.

Measures. Prediction performance is measured by the test accuracy for classification and the test $1-\mathrm{MAE}$ for regression. Fairness violation of a prediction head f on the representation $\mathbf{Z} = h_\phi(\mathbf{X})$ is measured through the disparity $\Delta_f(s) := |\mathbb{E}[f(\mathbf{Z}) \mid S = s] - \mathbb{E}[f(\mathbf{Z})]|$ at each sensitive value s. The prediction-level disparity $\mathrm{GDP}_\infty(f_\psi \circ h_\phi) = \sup_s \Delta_{f_\psi}(s)$ evaluates the prediction head f_ψ trained jointly with the encoder. Since a representation may later serve other prediction heads (Coston et al., 2021; Gupta et al., 2021; Jovanović et al., 2023; Luo et al., 2026), we also measure the largest disparity that a downstream prediction head can attain. For $c > 0$, let $\mathcal{F}_c := \{\sigma(\mathbf{w}^\top \mathbf{z} + b) : \|\mathbf{w}\|_2 \le c,\ b \in \mathbb{R}\}$ be the class of prediction heads with weight norm at most c, where $\sigma(t) := (1 + e^{-t})^{-1}$. We define the representation-level disparity over $\mathcal{F}_c$ as

$$\mathrm{GDP}_\infty(h_\phi; \mathcal{F}_c) := \sup_{f \in \mathcal{F}_c} \mathrm{GDP}_\infty(f \circ h_\phi) = \sup_{f \in \mathcal{F}_c} \sup_{s \in \mathcal{S}} \Delta_f(s). \tag{19}$$

We write $\widehat{\Delta}_f$ and $\widehat{\mathrm{GDP}}_\infty$ for the estimates of these quantities, which are defined in Appendix D.4.

Estimating the representation-level disparity. With the encoder held fixed, write $f_{\mathbf{w},b}(\mathbf{z}) := \sigma(\mathbf{w}^\top \mathbf{z} + b)$ and, on the training data, minimize

$$\widetilde{\mathcal{L}}(\mathbf{w}, b) := \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\mathbf{w},b}(\mathbf{Z}_i), Y_i) - \lambda_h \widehat{\mathrm{GDP}}_\infty(f_{\mathbf{w},b} \circ h_\phi) \tag{20}$$

over $\|\mathbf{w}\|_2 \le c$ and $b \in \mathbb{R}$, with $\lambda_h = 10^3$ so that the fairness regularization term dominates. Let $\hat{f}_c(\cdot) := f_{\hat{\mathbf{w}}_c, \hat{b}_c}(\cdot)$, where $(\hat{\mathbf{w}}_c, \hat{b}_c)$ minimizes $\widetilde{\mathcal{L}}(\mathbf{w}, b)$ subject to $\|\mathbf{w}\|_2 \le c$ and $b \in \mathbb{R}$. This head approximates the prediction head attaining the supremum in (19), and the representation-level disparity is estimated by $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_c) := \widehat{\mathrm{GDP}}_\infty(\hat{f}_c \circ h_\phi)$.

5.1 Simulation Study

Data generation. We generate an input $\mathbf{X} := (\mathbf{X}_{\mathrm{task}}, \mathbf{X}_{\mathrm{noise}}, X_{\mathrm{loc}})$ and an outcome Y as follows. We independently draw S on $[0, 1]$ with density $p(s) \propto 1 - 0.9\omega(s)$, $\mathbf{X}_{\mathrm{task}} \sim N(\mathbf{0}, \mathbf{I}_4)$, $\mathbf{X}_{\mathrm{noise}} \sim N(\mathbf{0}, \mathbf{I}_3)$ and $\varepsilon_1, \varepsilon_2 \sim N(0, 1)$ where $\omega(s) = \exp\{-(s - s_0)^2/(2\tau^2)\}$, and set

$$X_{\mathrm{loc}} := 3\omega(S) + 0.5\,\varepsilon_1, \qquad Y := \sigma\big(\boldsymbol{\upsilon}^\top \mathbf{X}_{\mathrm{task}} + \omega(S) + 0.3\,\varepsilon_2\big),$$

where $s_0 = 0.8$, $\tau = 0.043$ and $\boldsymbol{\upsilon} \in \mathbb{R}^4$ is a fixed vector. Since τ is small and so S affects Y only near s_0, the worst-case fairness violation can occur near s_0 where only a small proportion of observations fall in this narrow interval. Thus, EIPM may ignore unfairness around s_0. To confirm this conjecture, we compare supFR with FREM by varying λ in (14) and (5), respectively, and measure the representation-level disparity over $\mathcal{F}_1$. Details are provided in Appendix C.

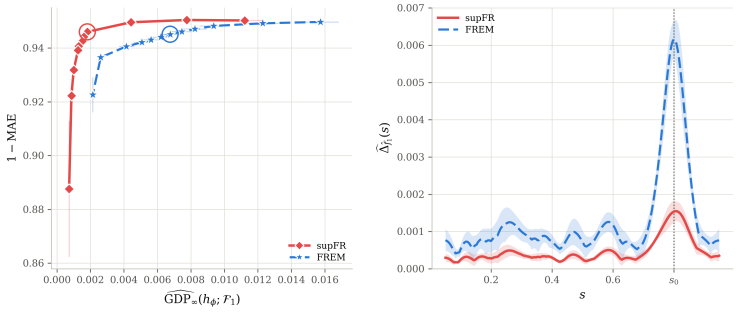


Figure 2: Left: prediction performance against $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$. Right: $\widehat{\Delta}_{\hat{f}_1}(s)$ for the two models circled in the left panel.

Results. In the left panel of Figure 2, supFR attains a lower representation-level disparity than FREM. The right panel shows $\widehat{\Delta}_{\hat{f}_1}(s)$ of the models at the circled points in the left panel. For FREM, it has a peak at s_0, whereas for supFR, it is markedly smaller there. This is the limitation of averaging the discrepancy over the sensitive attribute, shown in Figure 1. Since only a small proportion of the sensitive values lie in the narrow interval around s_0, even a large discrepancy in this interval has little effect on the EIPM in (2). In contrast, supFR penalizes the largest discrepancy over s in (9), and thus controls such a worst-case fairness violation.

5.2 Real data analysis

Datasets. We use three tabular datasets.

- Adult (Becker & Kohavi, 1996): Y is whether income exceeds \$50k and S is age.
- Crime (Redmond, 2002): Y is the violent crime rate and S is the proportion of Black residents.
- ACSIncome (Ding et al., 2021): Y is a binary income class in the 2018 American Community Survey and S is age.

Table 3 in Appendix D.1 gives further details, and Appendix D.5 reports results on the graph datasets Pokec-z and Pokec-n (Takac & Zabovsky, 2012).

Baselines. We compare supFR with seven baselines for fair representation learning (FRL) algorithms or for learning fair prediction models. Table 1 summarizes the type of sensitive attribute and the learning object of each method. The learning object indicates whether the method was proposed for fair representation learning or fair prediction. Appendix D.3 provides further details about the baseline methods. To apply LAFTR, MMD and sIPM-LFR to the continuous sensitive attribute S, we binarize the sensitive values at their median in the training data. For each method, we vary its fairness penalty parameter λ over a grid and train with five random seeds.

Table 1: Overview of fairness algorithms.

Method	Type of sensitive attribute	Learning object
supFR	continuous	fair representation
FREM (Kong et al., 2025)	continuous	fair representation
Reg-GDP (Jiang et al., 2022)	continuous or discrete	fair prediction
Reg-HGR (Mary et al., 2019)	continuous or discrete	fair prediction
ADV (Zhang et al., 2018)	continuous or discrete	fair prediction
LAFTR (Madras et al., 2018)	binary	fair representation
MMD (Deka & Sutherland, 2023)	binary	fair representation
sIPM-LFR (Kim et al., 2022)	binary	fair representation

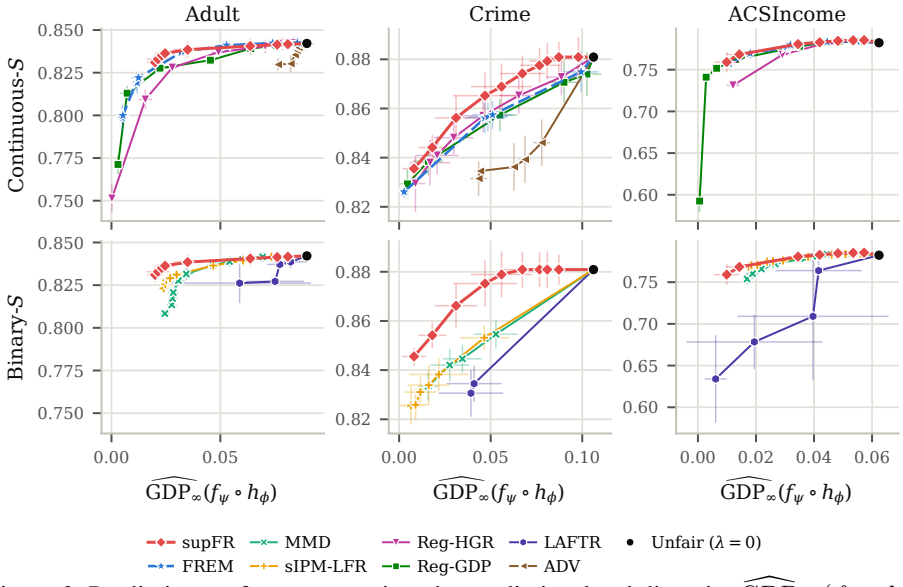


Figure 3: Prediction performance against the prediction-level disparity $\widehat{\mathrm{GDP}}_\infty(f_\psi \circ h_\phi)$.

Results. Figure 3 compares the methods by the prediction-level disparity and Figure 4 by the representation-level disparity over $\mathcal{F}_1$. supFR is similar to the other methods at the prediction level, whereas its representation-level disparity is markedly smaller than that of every other method on all three datasets. The prediction-level disparity concerns the trained head only, whereas the representation-level disparity concerns every head in $\mathcal{F}_1$.

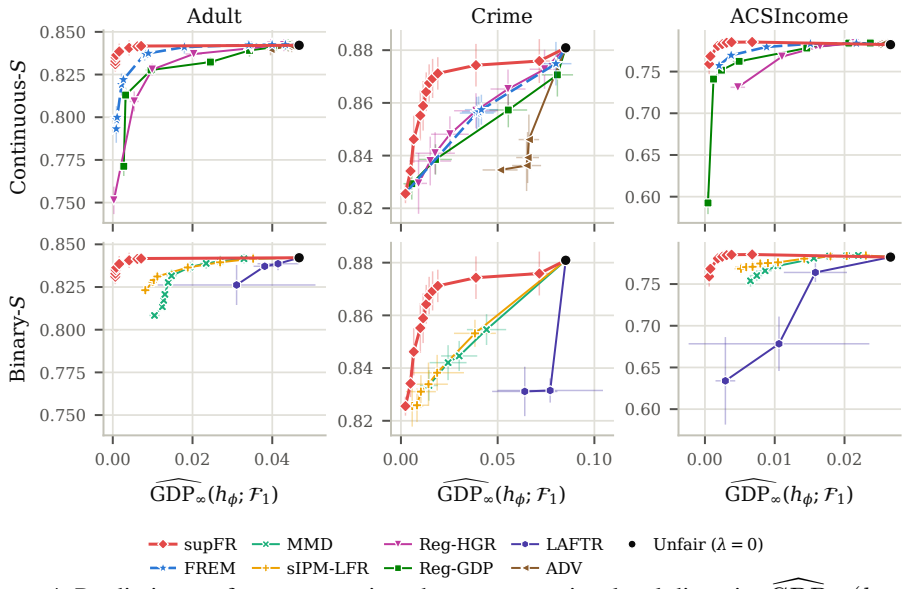


Figure 4: Prediction performance against the representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$.

The baselines do not directly control disparity at every sensitive value for all downstream heads, so the worst-case fairness violation may be large with other downstream heads. In contrast, supFR controls the largest ReLU-IPM between $\mathbb{P}_{\mathbf{Z}|S=s}$ and $\mathbb{P}_{\mathbf{Z}}$ over the sensitive domain, and Corollary 2 implies that the empirical supIPM yields a bound on $\Delta_f(s)$ over sensitive domain for every Hölder smooth head f. Figure 5 shows $\widehat{\Delta}_{\hat{f}_1}(s)$ on Adult for supFR and FREM with the test accuracies similar to that of the model trained without a fairness penalty. The disparity of supFR is small over the whole sensitive domain, whereas that of FREM is substantially larger for s near the left boundary and $s \approx 0.52$.

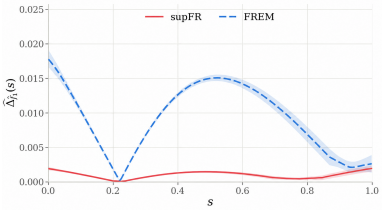


Figure 5: $\widehat{\Delta}_{\hat{f}_1}(s)$ of supFR and FREM at matched test accuracy.

Remark 2. *Unlike the right panel of Figure 2, FREM exhibits high $\widehat{\Delta}_{\hat{f}_1}(s)$ across the sensitive domain in Figure 5. A possible explanation is that the discriminator class used in FREM is more restrictive than the class of prediction heads whose fairness violation can be controlled by ReLU-IPM, as discussed in Section 3.2. We further examine the effect of the discriminator class in Appendix F.*

5.3 Additional results and ablation studies

We compare the methods under five further fairness measures and on the graph datasets in Appendix D.5, and the computational cost of supFR and FREM in Appendix E. Appendix F examines the effect of the discriminator class $\mathcal{V}$, and Appendix G shows the effects of the kernel K in (12) and of the number T of initial values of s.

6 Conclusion

We proposed supIPM, which measures the worst-case fairness violation of a representation for a continuous sensitive attribute. We then proposed supFR, which uses the empirical supIPM as a fairness penalty. We established a uniform convergence rate of the empirical supIPM, which yields a bound of the GDP_∞ for every Hölder smooth prediction head. On simulated and real data, supFR attains a markedly smaller representation-level disparity than existing methods at comparable prediction performance. Future work includes extending the fairness guarantees provided by supIPM to fairness criteria conditional on the outcome, such as equalized odds or equal opportunity.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing, for retrieval and discovery of related work, and for assistance in proving the mathematical claims in Appendix A; all AI-assisted content was reviewed by the authors.

ETHICS STATEMENT

This work aims to reduce fairness violations affecting under-represented populations defined by continuous sensitive attributes. Our guarantees concern the statistical fairness criteria and prediction-head classes specified in the paper, under the stated assumptions, and do not ensure the absence of all forms of discrimination. Practical applications should consider the appropriateness of the fairness criterion in the deployment context, potential biases in the data, and privacy when collecting and using sensitive attributes.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce our method and experiments in the Appendix. Appendices C and D provide the experimental settings, including the data-generating process of the simulation, the datasets, model architectures, training settings, baselines, and evaluation measures. Additional analyses of the computational cost and of the main design choices of supFR are provided in Appendices E, F, and G.

REFERENCES

Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair normalizing flows. *arXiv preprint arXiv:2106.05937*, 2021.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

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

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pp. 2144–2155. PMLR, 2021.

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

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 680–688, 2021.

Namrata Deka and Danica J Sutherland. Mmd-b-fair: Learning fair representations with statistical testing. In *International Conference on Artificial Intelligence and Statistics*, pp. 9564–9576. PMLR, 2023.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Xavier Gitiaux and Huzefa Rangwala. Learning smooth and fair representations. In *International conference on artificial intelligence and statistics*, pp. 253–261. PMLR, 2021.

Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 749–764. Springer, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7610–7619, 2021.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International conference on machine learning*, pp. 1929–1938. PMLR, 2018.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International conference on learning representations*, 2022.

Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, pp. 15401–15420. PMLR, 2023.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In *International conference on machine learning*, pp. 11074–11101. PMLR, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM review*, 45(3):385–482, 2003.

Insung Kong, Kunwoong Kim, and Yongdai Kim. Fair representation learning for continuous sensitive attributes using expectation of integral probability metrics. *IEEE transactions on pattern analysis and machine intelligence*, 47(5):3784–3795, 2025.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

Jihu Lee, Kunwoong Kim, Sehyun Park, Insung Kong, Dongyoon Yang, and Yongdai Kim. A fair bayesian inference through matched gibbs posterior. In *International Conference on Learning Representations*, volume 2026, pp. 60006–60037, 2026.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Yuhong Luo, Austin Hoag, Xintong Wang, Philip Thomas, and Przemyslaw Grabowicz. Fair representation learning with controllable high confidence guarantees via adversarial inference. *Advances in Neural Information Processing Systems*, 38:22503–22540, 2026.

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

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International conference on machine learning*, pp. 4382–4391. PMLR, 2019.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

Yuha Park, Kunwoong Kim, Insung Kong, and Yongdai Kim. Relu integral probability metric and its applications. *Neural Networks*, pp. 109589, 2026.

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C53W3X.

Brian C Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of statistics*, 2020.

Jinwon Sohn, Qifan Song, and Guang Lin. Fair supervised learning with a simple random sampler of sensitive attributes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1594–1602. PMLR, 2024.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International conference on artificial intelligence and statistics*, pp. 2164–2173. PMLR, 2019.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722. URL `https://doi.org/10.1214/12-EJS722`.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.

APPENDIX

A PROOFS OF THE MAIN RESULTS

This appendix states the assumptions of Section 4, proves four auxiliary lemmas, and then proves Theorem 1 and Corollary 2.

A.1 ASSUMPTIONS

We state the four conditions imposed in Section 4, with $\mathcal{S} = [0,1]$ as in Section 2.1 and $\mathcal{S}_n$, ϵ_n and γ_n as in Section 4.

Assumption 1 (Positive density)**.** *$\mathbb{P}_S$ has a density p_S with respect to the Lebesgue measure* Leb *on $\mathcal{S} = [0,1]$, and $0 < c_S \le p_S(s) \le C_S < \infty$ for every $s \in \mathcal{S}$.*

Let ν be a σ-finite measure on $\mathcal{X}$ and $p(\mathbf{x}, s)$ a joint density of $(\mathbf{X}, S)$ with respect to $\nu \otimes \mathrm{Leb}$ with $\int_{\mathcal{X}} p(\mathbf{x}, s)\,\mathrm{d}\nu(\mathbf{x}) = p_S(s)$ for every $s \in \mathcal{S}$. For a bounded measurable $g : \mathcal{X} \to \mathbb{R}$, this joint density fixes the version of the conditional expectation given $S = s$ used throughout the paper,

$$\mathbb{E}[g(\mathbf{X}) \mid S = s] := \frac{1}{p_S(s)} \int_{\mathcal{X}} g(\mathbf{x})\, p(\mathbf{x}, s)\,\mathrm{d}\nu(\mathbf{x}), \qquad s \in \mathcal{S}, \tag{21}$$

which Equations (9) and (11) require at every $s \in \mathcal{S}$, not only for $\mathbb{P}_S$-almost every s.

Assumption 2 (Smooth joint density)**.** *The sample $\mathcal{D}$ consists of independent copies of $(\mathbf{X}, S, Y)$. For ν-almost every $\mathbf{x}$ the map $s \mapsto p(\mathbf{x}, s)$ is twice continuously differentiable on $\mathcal{S}$, and there exist $M_1, M_2 \in L_1(\nu)$ such that $\sup_{s \in \mathcal{S}} \left|\partial_s^j p(\mathbf{x}, s)\right| \le M_j(\mathbf{x})$ for $j \in \{1, 2\}$.*

Assumption 2 imposes no smoothness in $\mathbf{x}$. The integrable envelopes M_1 and M_2 justify differentiation under the integral sign with respect to s.

Assumption 3 (Kernel and bandwidth)**.** *$K_\gamma(s', s) = \gamma^{-1} K((s' - s)/\gamma)$ for a bounded, symmetric and Lipschitz function $K : \mathbb{R} \to [0, \infty)$ with $\int K = 1$ and $\int t^2 K(t)\,\mathrm{d}t < \infty$. With $T_K(r) := \int_{|t| > r} K(t)\,\mathrm{d}t$ the tail mass of K, the bandwidth satisfies $\gamma_n \to 0$, $0 < \gamma_n \le \min\{\epsilon_n, 1/2\}$ and $T_K(\epsilon_n/\gamma_n) \le \gamma_n^2$.*

The tail condition of Assumption 3 bounds the kernel mass beyond $\mathcal{S}$ by γ_n^2 at every $s \in \mathcal{S}_n$. For a kernel supported on $[-1, 1]$ it holds automatically, because $\epsilon_n/\gamma_n \ge 1$ and hence $T_K(\epsilon_n/\gamma_n) = 0$. For the Gaussian kernel it holds whenever $\epsilon_n \asymp 1/\log n$ and γ_n decays polynomially.

Assumption 4 (Encoder class)**.** *$\mathcal{E}_n$ is a pointwise measurable class (van der Vaart & Wellner, 1996, Example 2.3.4) of measurable encoders $h : \mathcal{X} \to \mathbb{B}^d$. For encoders h and h' let*

$$\|h - h'\|_\infty := \sup_{\mathbf{x} \in \mathcal{X}} \|h(\mathbf{x}) - h'(\mathbf{x})\|_2, \tag{22}$$

and let $N(\eta, \mathcal{F}, \|\cdot\|_\infty)$ be the smallest number of $\|\cdot\|_\infty$-balls of radius η with centers in $\mathcal{F}$ that cover a class $\mathcal{F}$. There exist $A_n \ge e$ and $\mathfrak{v}_n \ge 1$ such that

$$N(\eta, \mathcal{E}_n, \|\cdot\|_\infty) \le \left(\frac{A_n}{\eta}\right)^{\mathfrak{v}_n}, \qquad 0 < \eta \le 1. \tag{23}$$

Assumption 4 holds on a bounded $\mathcal{X}$ for ReLU networks with bounded parameters followed by the normalization of Appendix B.1, which is 1-Lipschitz because its Jacobian is symmetric with eigenvalues in $(0, 1]$, and therefore does not increase covering numbers. For such networks, $\mathfrak{v}_n$ is bounded by a multiple of the number of parameters and $\log A_n$ by a multiple of the depth times the logarithm of the width (Schmidt-Hieber, 2020).

A.2 NOTATION AND AUXILIARY INEQUALITIES

Throughout the appendix $\mathbb{E}$ denotes expectation under the joint distribution of $(\mathbf{X}, S)$ and $\mathbb{E}_n$ the empirical average, $\mathbb{E}_n\varphi := n^{-1}\sum_{i=1}^n \varphi(\mathbf{X}_i, S_i)$. We abbreviate $\gamma := \gamma_n$ and $\epsilon := \epsilon_n$,

and for notational convenience we write $D_n(h) := \mathrm{supIPM}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X}); S)$ and $\widehat{D}_n(h) := \widehat{\mathrm{supIPM}}^{\gamma_n}_{\mathcal{V}_\mathrm{R},n}(h(\mathbf{X}); S)$ for the quantities in (15). We write C for a finite positive constant depending only on c_S, C_S, $\|M_1\|_{L_1(\nu)}$, $\|M_2\|_{L_1(\nu)}$ and K, whose value may change from one occurrence to the next.

The encoder class $\mathcal{E}_n$ and the ReLU discriminator class $\mathcal{V}_\mathrm{R}$ induce the class

$$\mathcal{G}_n := \{\mathbf{x} \mapsto v(h(\mathbf{x})) : h \in \mathcal{E}_n,\ v \in \mathcal{V}_\mathrm{R}\}. \tag{24}$$

Every $g \in \mathcal{G}_n$ satisfies $0 \le g \le 2$, because h takes values in $\mathbb{B}^d$ and $0 \le (\theta^\top \mathbf{z} + \mu)_+ \le \|\theta\|_2 \|\mathbf{z}\|_2 + |\mu| \le 2$ for $\mathbf{z} \in \mathbb{B}^d$, $\theta \in \mathbb{S}^{d-1}$ and $\mu \in [-1, 1]$. For a measurable $g : \mathcal{X} \to [0, 2]$ and $s \in \mathcal{S}$ let

$$\begin{aligned} a_g(s) &:= \int_\mathcal{X} g(\mathbf{x}) p(\mathbf{x}, s)\, \mathrm{d}\nu(\mathbf{x}), & m_g(s) &:= \frac{a_g(s)}{p_S(s)} = \mathbb{E}[g(\mathbf{X}) \mid S = s], \\ b_{g,\gamma}(s) &:= \mathbb{E}\{K_\gamma(S, s) g(\mathbf{X})\}, & q_\gamma(s) &:= \mathbb{E} K_\gamma(S, s), \\ m_{g,\gamma}(s) &:= \frac{b_{g,\gamma}(s)}{q_\gamma(s)}, & \widehat{m}_{g,\gamma}(s) &:= \frac{\mathbb{E}_n\{K_\gamma(S, s) g(\mathbf{X})\}}{\mathbb{E}_n K_\gamma(S, s)}, \end{aligned} \tag{25}$$

where m_g is the version of the conditional expectation fixed in (21). The quantity $\widehat{m}_{g,\gamma}$ is the Nadaraya–Watson estimator of m_g, and $m_{g,\gamma}$ is its population counterpart. By the definition of $\widehat{w}_\gamma(i; s)$ in Section 3.3, $\widehat{m}_{g,\gamma}(s) = \sum_{i=1}^n \widehat{w}_\gamma(i; s) g(\mathbf{X}_i)$ whenever $\mathbb{E}_n K_\gamma(S, s) > 0$. Hence, for every $h \in \mathcal{E}_n$ and whenever $\inf_{s \in \mathcal{S}_n} \mathbb{E}_n K_\gamma(S, s) > 0$,

$$D_n(h) = \sup_{g,s} |m_g(s) - \mathbb{E} g|, \qquad \widehat{D}_n(h) = \sup_{g,s} |\widehat{m}_{g,\gamma}(s) - \mathbb{E}_n g|, \tag{26}$$

where the suprema are taken over $g = v \circ h$ with $v \in \mathcal{V}_\mathrm{R}$ and over $s \in \mathcal{S}_n$. Because g, m_g, $m_{g,\gamma}$ and $\widehat{m}_{g,\gamma}$ all take values in $[0, 2]$, $D_n(h)$ and $\widehat{D}_n(h)$ lie in $[0, 2]$.

We use Bernstein's inequality in the following form (Boucheron et al., 2013, Chapter 2). If $W_1, \dots, W_n$ are independent copies of a random variable W with $0 \le W \le U$ and $\mathrm{Var}(W) \le \sigma^2$, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n W_i - \mathbb{E} W\right| \ge u\right) \le 2 \exp\left(-\frac{n u^2}{2\sigma^2 + 2Uu/3}\right), \qquad u > 0. \tag{27}$$

We also use the inequality

$$\left|\sup_{t \in \mathcal{T}} |F_1(t)| - \sup_{t \in \mathcal{T}} |F_2(t)|\right| \le \sup_{t \in \mathcal{T}} |F_1(t) - F_2(t)|. \tag{28}$$

This inequality holds for any bounded functions $F_1, F_2 : \mathcal{T} \to \mathbb{R}$. Indeed, the triangle inequality gives

$$\sup_{t \in \mathcal{T}} |F_1(t)| \le \sup_{t \in \mathcal{T}} |F_2(t)| + \sup_{t \in \mathcal{T}} |F_1(t) - F_2(t)|.$$

Interchanging F_1 and F_2 gives the reverse inequality.

A.3 Auxiliary lemmas

Lemma 3 (Lipschitz continuity and covering number bound)**.** *The ReLU discriminators and the induced class $\mathcal{G}_n$ satisfy the following properties.*
(i) *For $\mathbf{z}, \mathbf{z}' \in \mathbb{B}^d$, $\theta, \theta' \in \mathbb{S}^{d-1}$ and $\mu, \mu' \in [-1, 1]$,*

$$|v_{\theta,\mu}(\mathbf{z}) - v_{\theta',\mu'}(\mathbf{z}')| \le \|\mathbf{z} - \mathbf{z}'\|_2 + \|\theta - \theta'\|_2 + |\mu - \mu'|.$$

(ii) *Under Assumption 4,*

$$N(\eta, \mathcal{G}_n, \|\cdot\|_\infty) \le \left(\frac{9 A_n}{\eta}\right)^{\mathfrak{v}_n + d + 1}, \qquad 0 < \eta \le 1.$$

Proof. (i) The map $t \mapsto (t)_+$ is 1-Lipschitz on $\mathbb{R}$, so

$$|v_{\theta,\mu}(\mathbf{z}) - v_{\theta',\mu'}(\mathbf{z}')| \le \left|\theta^\top \mathbf{z} + \mu - \theta'^\top \mathbf{z}' - \mu'\right| \le \left|\theta^\top (\mathbf{z} - \mathbf{z}')\right| + \left|(\theta - \theta')^\top \mathbf{z}'\right| + |\mu - \mu'|,$$

and the Cauchy–Schwarz inequality together with $\|\theta\|_2 = 1$ and $\|\mathbf{z}'\|_2 \le 1$ bounds the first two terms by $\|\mathbf{z} - \mathbf{z}'\|_2$ and $\|\theta - \theta'\|_2$.

(ii) Let $0 < \eta \le 1$. Since $h(\mathbf{x}), h'(\mathbf{x}) \in \mathbb{B}^d$ for $h, h' \in \mathcal{E}_n$ by Assumption 4, part (i) gives

$$\|v_{\theta,\mu} \circ h - v_{\theta',\mu'} \circ h'\|_\infty \le \|h - h'\|_\infty + \|\theta - \theta'\|_2 + |\mu - \mu'|. \tag{29}$$

Take an $(\eta/2)$-net of $\mathcal{E}_n$ in $\|\cdot\|_\infty$, of cardinality at most $(2A_n/\eta)^{\mathfrak{v}_n}$ by Assumption 4. Next take an $(\eta/4)$-net of $\mathbb{S}^{d-1}$ in $\|\cdot\|_2$, of cardinality at most $(1 + 8/\eta)^d \le (9/\eta)^d$ by the standard volumetric bound for subsets of the unit ball. Finally take an $(\eta/4)$-net of $[-1, 1]$, of cardinality at most $\lceil 4/\eta \rceil \le 5/\eta$. By (29), the functions $v_{\theta,\mu} \circ h$ with h, θ and μ taken from these three nets form an η-net of $\mathcal{G}_n$, and its cardinality is at most

$$\left(\frac{2A_n}{\eta}\right)^{\mathfrak{v}_n} \left(\frac{9}{\eta}\right)^d \frac{5}{\eta} \le \left(\frac{9A_n}{\eta}\right)^{\mathfrak{v}_n + d + 1},$$

because $A_n \ge e$. □

Lemma 4. *Let Assumptions 1 and 2 hold and let $g : \mathcal{X} \to [0, 2]$ be measurable. Then a_g and p_S are twice continuously differentiable on $\mathcal{S}$, with*

$$a_g^{(j)}(s) = \int_{\mathcal{X}} g(\mathbf{x})\, \partial_s^j p(\mathbf{x}, s)\, \mathrm{d}\nu(\mathbf{x}), \qquad \|a_g^{(j)}\|_\infty \le 2\|M_j\|_{L_1(\nu)}, \qquad \|p_S^{(j)}\|_\infty \le \|M_j\|_{L_1(\nu)}$$

for $j \in \{1, 2\}$. Moreover $0 \le a_g \le 2p_S \le 2C_S$ on $\mathcal{S}$, and consequently $0 \le m_g \le 2$.

Proof. Fix $s \in \mathcal{S}$ and let $t \ne 0$ be such that $s + t \in \mathcal{S}$. For ν-almost every $\mathbf{x}$ the map $s \mapsto p(\mathbf{x}, s)$ is differentiable on $\mathcal{S}$, so by the mean value theorem there is $\xi_{\mathbf{x},t}$ between s and $s + t$ with

$$\left| g(\mathbf{x}) \frac{p(\mathbf{x}, s + t) - p(\mathbf{x}, s)}{t} \right| = |g(\mathbf{x})\, \partial_s p(\mathbf{x}, \xi_{\mathbf{x},t})| \le 2M_1(\mathbf{x}).$$

Since $M_1 \in L_1(\nu)$, dominated convergence applies as $t \to 0$ and yields $a_g'(s) = \int g\, \partial_s p\, \mathrm{d}\nu$. Repeating the argument with $\partial_s p$ in place of p and M_2 in place of M_1 gives $a_g''(s) = \int g\, \partial_s^2 p\, \mathrm{d}\nu$. Continuity of a_g'' follows from the continuity of $s \mapsto \partial_s^2 p(\mathbf{x}, s)$ and a further application of dominated convergence with envelope $2M_2$. The stated bounds on $\|a_g^{(j)}\|_\infty$ are immediate from $0 \le g \le 2$, and the assertions for p_S follow by taking $g \equiv 1$, for which the envelope is M_j rather than $2M_j$. Finally $0 \le a_g(s) \le 2 \int p(\mathbf{x}, s)\, \mathrm{d}\nu(\mathbf{x}) = 2p_S(s) \le 2C_S$ by Assumption 1, and dividing by $p_S(s) > 0$ gives $0 \le m_g \le 2$. □

Lemma 5. *Let Assumptions 1–3 hold. Then, for every n and every $s \in \mathcal{S}_n$,*

$$\tfrac{3}{4} c_S \;\le\; q_\gamma(s) \;\le\; C_S, \qquad |q_\gamma(s) - p_S(s)| \le C\gamma^2, \tag{30}$$

and, uniformly over measurable $g : \mathcal{X} \to [0, 2]$,

$$\sup_{s \in \mathcal{S}_n} |b_{g,\gamma}(s) - a_g(s)| \le C\gamma^2, \qquad \sup_{s \in \mathcal{S}_n} |m_{g,\gamma}(s) - m_g(s)| \le C\gamma^2. \tag{31}$$

Proof. Extend $p(\mathbf{x}, \cdot)$ and p_S by zero outside $\mathcal{S}$, so that a_g and p_S are defined on all of $\mathbb{R}$ and vanish off $\mathcal{S}$. By Assumption 3 and the substitution $s' = s + \gamma t$,

$$b_{g,\gamma}(s) = \int_{\mathbb{R}} K(t)\, a_g(s + \gamma t)\, \mathrm{d}t, \qquad q_\gamma(s) = \int_{\mathbb{R}} K(t)\, p_S(s + \gamma t)\, \mathrm{d}t. \tag{32}$$

Fix $s \in \mathcal{S}_n$ and set $I_n := [-\epsilon/\gamma, \epsilon/\gamma]$. For every $t \in I_n$ we have $s + \gamma t \in [s - \epsilon, s + \epsilon] \subseteq \mathcal{S}$. Write $\kappa_K := \int_{\mathbb{R}} t^2 K(t)\, \mathrm{d}t$, which is finite by Assumption 3.

Since $m_g = a_g / p_S$ and $m_{g,\gamma} = b_{g,\gamma}/q_\gamma$, the difference to be bounded decomposes as

$$\begin{aligned} m_{g,\gamma}(s) - m_g(s) &= \frac{b_{g,\gamma}(s)}{q_\gamma(s)} - \frac{a_g(s)}{p_S(s)} = \frac{b_{g,\gamma}(s) - a_g(s)}{q_\gamma(s)} + \frac{a_g(s)}{p_S(s)} \cdot \frac{p_S(s) - q_\gamma(s)}{q_\gamma(s)} \\ &= \frac{b_{g,\gamma}(s) - a_g(s)}{q_\gamma(s)} + m_g(s)\, \frac{p_S(s) - q_\gamma(s)}{q_\gamma(s)}. \end{aligned} \tag{33}$$

Step (i) bounds the denominator $q_\gamma(s)$ from above and below, Step (ii) bounds the two smoothing biases $b_{g,\gamma}(s) - a_g(s)$ and $q_\gamma(s) - p_S(s)$, and Step (iii) combines the bounds through (33).

(i) The denominator. By Lemma 4 and Assumption 1, $0 \le p_S \le C_S$ everywhere after the extension, so $q_\gamma(s) \le C_S \int K = C_S$. In addition, $p_S \ge c_S$ on $\mathcal{S}$ gives

$$q_\gamma(s) \;\ge\; \int_{I_n} K(t) p_S(s+\gamma t)\,\mathrm{d}t \;\ge\; c_S \int_{I_n} K(t)\,\mathrm{d}t \;=\; c_S\{1 - T_K(\epsilon/\gamma)\} \;\ge\; c_S(1-\gamma^2) \;\ge\; \tfrac{3}{4}c_S,$$

where the third inequality is the tail condition $T_K(\epsilon/\gamma) \le \gamma^2$ of Assumption 3 and the last uses $\gamma \le 1/2$. This proves the first bound in (30).

(ii) The numerator. Write $\int_{\mathbb{R}} = \int_{I_n} + \int_{I_n^c}$ in the first identity of (32). Since $0 \le a_g \le 2C_S$ by Lemma 4,

$$\left| \int_{I_n^c} K(t) a_g(s+\gamma t)\,\mathrm{d}t \right| \le 2C_S \int_{I_n^c} K(t)\,\mathrm{d}t = 2C_S T_K(\epsilon/\gamma) \le 2C_S\gamma^2. \tag{34}$$

On I_n the segment joining s and $s+\gamma t$ is contained in $\mathcal{S}$, so Taylor's theorem and the bound $\|a_g''\|_\infty \le 2\|M_2\|_{L_1(\nu)}$ of Lemma 4 give

$$\left| a_g(s+\gamma t) - a_g(s) - \gamma t\, a_g'(s) \right| \le \frac{\gamma^2 t^2}{2} \|a_g''\|_\infty \le \gamma^2 t^2 \|M_2\|_{L_1(\nu)}, \qquad t \in I_n.$$

Multiply by $K(t)$ and integrate over I_n. Because K and I_n are symmetric about the origin, $\int_{I_n} tK(t)\,\mathrm{d}t = 0$, and $\int_{I_n} K(t)\,\mathrm{d}t = 1 - T_K(\epsilon/\gamma)$, so

$$\begin{aligned}
\left| \int_{I_n} K(t) a_g(s+\gamma t)\,\mathrm{d}t - a_g(s) \right| &\le \left| \int_{I_n} K(t)\{a_g(s+\gamma t) - a_g(s) - \gamma t a_g'(s)\}\,\mathrm{d}t \right| \\
&\quad + a_g(s) \left| \int_{I_n} K(t)\,\mathrm{d}t - 1 \right| \\
&\le \gamma^2 \|M_2\|_{L_1(\nu)} \kappa_K + a_g(s) T_K(\epsilon/\gamma).
\end{aligned}$$

Adding (34), and using $T_K(\epsilon/\gamma) \le \gamma^2$ and $a_g \le 2C_S$, we obtain

$$\begin{aligned}
|b_{g,\gamma}(s) - a_g(s)| &\le \left| \int_{I_n} K(t) a_g(s+\gamma t)\,\mathrm{d}t - a_g(s) \right| + \left| \int_{I_n^c} K(t) a_g(s+\gamma t)\,\mathrm{d}t \right| \\
&\le \left(4C_S + \|M_2\|_{L_1(\nu)}\kappa_K\right)\gamma^2,
\end{aligned} \tag{35}$$

which is the first bound in (31). For $g \equiv 1$, $a_g = p_S \le C_S$ and $\|p_S''\|_\infty \le \|M_2\|_{L_1(\nu)}$ by Lemma 4, so the tail term, the term $p_S(s)T_K(\epsilon/\gamma)$ and the Taylor remainder are at most $C_S\gamma^2$, $C_S\gamma^2$ and $\frac{1}{2}\gamma^2\|M_2\|_{L_1(\nu)}\kappa_K$. The same argument then gives

$$|q_\gamma(s) - p_S(s)| \le \left(2C_S + \tfrac{1}{2}\|M_2\|_{L_1(\nu)}\kappa_K\right)\gamma^2, \tag{36}$$

which is the second bound in (30).

(iii) The ratio. Taking absolute values in (33) and inserting $q_\gamma(s) \ge 3c_S/4$ from Step (i), (35) and (36) from Step (ii), and $0 \le m_g \le 2$ from Lemma 4,

$$\begin{aligned}
|m_{g,\gamma}(s) - m_g(s)| &\le \frac{|b_{g,\gamma}(s) - a_g(s)|}{q_\gamma(s)} + m_g(s)\,\frac{|p_S(s) - q_\gamma(s)|}{q_\gamma(s)} \\
&\le \frac{4}{3c_S}\left\{\left(4C_S + \|M_2\|_{L_1(\nu)}\kappa_K\right) + 2\left(2C_S + \tfrac{1}{2}\|M_2\|_{L_1(\nu)}\kappa_K\right)\right\}\gamma^2 \\
&= \frac{8\left(4C_S + \|M_2\|_{L_1(\nu)}\kappa_K\right)}{3c_S}\,\gamma^2,
\end{aligned}$$

which is the second bound in (31). None of the constants depends on g or on s. □

Lemma 6 (Uniform stochastic error)**.** *Let Assumptions 1–4 hold. There exist constants $C_*, c > 0$, depending only on c_S, C_S and K, such that for every $\delta \in (0,1)$ and every n with $\Gamma_{n,\delta} \le cn\gamma$ the event*

$$\Omega_n := \left\{ \sup_{g\in\mathcal{G}_n}\sup_{s\in\mathcal{S}_n} |(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s)g(\mathbf{X})\}| \le C_*\Lambda_n,\ \sup_{s\in\mathcal{S}_n} |(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s)| \le C_*\Lambda_n,\right.$$
$$\left. \sup_{g\in\mathcal{G}_n} |(\mathbb{E}_n - \mathbb{E})g| \le C_*\Lambda_n \right\} \tag{37}$$

satisfies $\mathbb{P}(\Omega_n) \ge 1-\delta$, where $\Lambda_n := \sqrt{\frac{\Gamma_{n,\delta}}{n\gamma}}$. Moreover, on the event Ω_n,

$$\inf_{s\in\mathcal{S}_n} \mathbb{E}_n K_\gamma(S,s) \ge c_S/2 \tag{38}$$

holds.

Proof. The constants are determined in the following order. Step 1 fixes a constant $C_0 \ge 1$ depending only on K, Step 2 fixes $C_* \ge 1$ depending only on c_S, C_S and K, and c is then defined by

$$c := \min\left\{\frac{1}{C_*^2},\ \frac{c_S^2}{16C_*^2}\right\}, \qquad \text{so that} \qquad c \le 1, \quad C_*\sqrt{c} \le 1 \quad \text{and} \quad C_*\sqrt{c} \le \frac{c_S}{4}. \tag{39}$$

Neither C_0 nor C_* depends on c, and only the three consequences listed in (39) are used below.

Since $\mathfrak{v}_n \ge 1$, $d \ge 1$, $A_n \ge e$ and $\gamma \le 1$, the definition in (16) gives $\Gamma_{n,\delta} \ge (\mathfrak{v}_n + d)\log A_n \ge 2$, and in particular $\Gamma_{n,\delta} \ge 1$. Combining this with the hypothesis $\Gamma_{n,\delta} \le cn\gamma$ and with $\gamma \le 1$ and $n \ge 1$,

$$\gamma \ge \frac{\Gamma_{n,\delta}}{cn} \ge \frac{1}{cn}, \qquad \Lambda_n = \sqrt{\frac{\Gamma_{n,\delta}}{n\gamma}} \ge \sqrt{\frac{1}{n\gamma}} \ge \frac{1}{\sqrt{n}} \ge \frac{1}{n}, \qquad \Lambda_n \le \sqrt{\frac{cn\gamma}{n\gamma}} = \sqrt{c}. \tag{40}$$

Step 1: the numerator class. Let $u := C_*\Lambda_n$, where $C_* \ge 1$ is the constant to be fixed in Step 2. By (40) and the consequence $C_*\sqrt{c} \le 1$ of (39),

$$\frac{1}{n} \le \Lambda_n \le u \le C_*\sqrt{c} \le 1. \tag{41}$$

Let $L_K > 0$ be a Lipschitz constant of K and put

$$\zeta_{\mathcal{G}} := \min\left\{\frac{\gamma u}{8\|K\|_\infty},\ 1\right\}, \qquad \zeta_{\mathcal{S}} := \frac{\gamma^2 u}{16 L_K}.$$

Since $0 < \zeta_{\mathcal{G}} \le 1$, Lemma 3(ii) provides a $\zeta_{\mathcal{G}}$-net $g_1,\dots,g_{N_1}$ of $\mathcal{G}_n$ in $\|\cdot\|_\infty$ with $0 \le g_k \le 2$. Let $s_1,\dots,s_{N_2}$ be a $\zeta_{\mathcal{S}}$-net of $\mathcal{S}_n$ in $|\cdot|$. Given $g \in \mathcal{G}_n$ and $s \in \mathcal{S}_n$, choose k and j with $\|g - g_k\|_\infty \le \zeta_{\mathcal{G}}$ and $|s - s_j| \le \zeta_{\mathcal{S}}$. Since K is L_K-Lipschitz, for every $s' \in \mathcal{S}$,

$$\begin{aligned} |K_\gamma(s',s) - K_\gamma(s',s_j)| &= \frac{1}{\gamma}\left|K\left(\frac{s'-s}{\gamma}\right) - K\left(\frac{s'-s_j}{\gamma}\right)\right| \\ &\le \frac{L_K}{\gamma}\cdot\frac{|s-s_j|}{\gamma} \le \frac{L_K}{\gamma^2}\zeta_{\mathcal{S}} = \frac{u}{16}. \end{aligned} \tag{42}$$

Hence, using $K_\gamma \le \|K\|_\infty/\gamma$, $0 \le g_k \le 2$ and $\zeta_{\mathcal{G}} \le \gamma u/(8\|K\|_\infty)$,

$$\begin{aligned} &\sup_{\mathbf{x}\in\mathcal{X},\, s'\in\mathcal{S}} |K_\gamma(s',s)g(\mathbf{x}) - K_\gamma(s',s_j)g_k(\mathbf{x})| \\ &\quad\le \sup_{s'\in\mathcal{S}} K_\gamma(s',s)\,\|g-g_k\|_\infty + \sup_{s'\in\mathcal{S}} |K_\gamma(s',s) - K_\gamma(s',s_j)|\,\|g_k\|_\infty \\ &\quad\le \frac{\|K\|_\infty}{\gamma}\zeta_{\mathcal{G}} + \frac{2u}{16} \le \frac{u}{8} + \frac{u}{8} = \frac{u}{4}. \end{aligned} \tag{43}$$

Applying $\mathbb{E}_n$ and $\mathbb{E}$ to the bound in (43) gives

$$|(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s)g(\mathbf{X})\}| \le |(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s_j)g_k(\mathbf{X})\}| + \frac{u}{4} + \frac{u}{4},$$

so if the left-hand side exceeds u, then $|(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s_j)g_k(\mathbf{X})\}| > u/2$. Therefore, with the suprema running over $g \in \mathcal{G}_n$ and $s \in \mathcal{S}_n$,

$$\mathbb{P}\left(\sup_{g,s} |(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s)g(\mathbf{X})\}| > u\right) \le \sum_{k=1}^{N_1}\sum_{j=1}^{N_2} \mathbb{P}\left(|(\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s_j)g_k(\mathbf{X})\}| > \tfrac{u}{2}\right). \tag{44}$$

The variable $W := K_\gamma(S,s_j)g_k(\mathbf{X})$ satisfies $0 \le W \le U := 2\|K\|_\infty/\gamma$ and, by $K_\gamma \le \|K\|_\infty/\gamma$, $g_k \le 2$ and $q_\gamma \le C_S$ from Lemma 5,

$$\mathrm{Var}(W) \le \mathbb{E}W^2 = \mathbb{E}\{K_\gamma(S,s_j)^2 g_k(\mathbf{X})^2\} \le \frac{4\|K\|_\infty}{\gamma}\,\mathbb{E}K_\gamma(S,s_j) \le \frac{4\|K\|_\infty C_S}{\gamma} =: \sigma^2.$$

By (27) with $u/2$ in place of u, each summand in (44) is at most

$$2\exp\left\{-\frac{n(u/2)^2}{2\sigma^2 + 2U(u/2)/3}\right\} = 2\exp\left\{-\frac{nu^2}{8\sigma^2 + 4Uu/3}\right\},$$

so the right-hand side of (44) is at most $2N_1N_2\exp\{-nu^2/(8\sigma^2+4Uu/3)\}$, which is at most $\delta/3$ if and only if

$$\frac{nu^2}{8\sigma^2 + 4Uu/3} \;\ge\; \Xi, \qquad \Xi := \log\frac{6N_1N_2}{\delta}. \tag{45}$$

If $nu^2 \ge 16\Xi\sigma^2$ and $nu^2 \ge 8\Xi Uu/3$, then $8\sigma^2 \le nu^2/(2\Xi)$ and $4Uu/3 \le nu^2/(2\Xi)$, so $8\sigma^2 + 4Uu/3 \le nu^2/\Xi$ and (45) holds. With $\sigma^2 = 4\|K\|_\infty C_S/\gamma$ and $U = 2\|K\|_\infty/\gamma$, these two conditions read

$$u \;\ge\; 4\sigma\sqrt{\frac{\Xi}{n}} = 8\sqrt{\frac{\|K\|_\infty C_S \Xi}{n\gamma}} \qquad \text{and} \qquad u \;\ge\; \frac{8\Xi U}{3n} = \frac{16\|K\|_\infty \Xi}{3n\gamma}. \tag{46}$$

It remains to bound Ξ. By Lemma 3(ii) with $\eta = \zeta_{\mathcal{G}}$, and since $1/\zeta_{\mathcal{G}} = \max\{8\|K\|_\infty/(\gamma u), 1\} \le (8\|K\|_\infty + 1)/(\gamma u)$ by $\gamma u \le 1$,

$$\log N_1 \le (\mathfrak{v}_n + d + 1)\log\frac{9A_n}{\zeta_{\mathcal{G}}} \le (\mathfrak{v}_n + d + 1)\log\frac{9(8\|K\|_\infty + 1)A_n}{\gamma u}.$$

The interval $\mathcal{S}_n \subseteq [0,1]$ admits a $\zeta_{\mathcal{S}}$-net with at most $1 + 1/(2\zeta_{\mathcal{S}})$ points, so

$$\log N_2 \le \log\left(1 + \frac{1}{2\zeta_{\mathcal{S}}}\right) = \log\left(1 + \frac{8L_K}{\gamma^2 u}\right).$$

We now use $u \ge n^{-1}$ from (41), which gives $\log(1/u) \le \log n$. Since $A_n \ge e$ and $\gamma \le 1$, $\log(A_n n/\gamma) = \log A_n + \log n + \log(1/\gamma) \ge 1$, and since $\gamma^2 u \le 1$, $1 + 8L_K/(\gamma^2 u) \le (1 + 8L_K)/(\gamma^2 u)$. Hence

$$\begin{aligned}
\log\frac{9(8\|K\|_\infty + 1)A_n}{\gamma u} &\le \log\{9(8\|K\|_\infty + 1)\} + \log A_n + \log\frac{1}{\gamma} + \log n \\
&\le [1 + \log\{9(8\|K\|_\infty + 1)\}]\log\frac{A_n n}{\gamma}, \\
\log\left(1 + \frac{8L_K}{\gamma^2 u}\right) &\le \log(1 + 8L_K) + 2\log\frac{1}{\gamma} + \log n \\
&\le [2 + \log(1 + 8L_K)]\log\frac{A_n n}{\gamma}.
\end{aligned}$$

Since $\mathfrak{v}_n + d + 1 \le 2(\mathfrak{v}_n + d)$ and $\mathfrak{v}_n + d \ge 1$, we conclude that $\Xi = \log 6 + \log N_1 + \log N_2 + \log(1/\delta)$ satisfies

$$\begin{aligned}
\Xi &\le C_0\left\{(\mathfrak{v}_n + d)\log\frac{A_n n}{\gamma} + \log\frac{1}{\delta}\right\} = C_0\,\Gamma_{n,\delta}, \\
C_0 &:= 4 + 2\log\{36(8\|K\|_\infty + 1)\} + \log(1 + 8L_K) + \log 6,
\end{aligned} \tag{47}$$

where $C_0 \geq 1$ depends only on K, and where the factor 36 in place of 9 is chosen so that the same C_0 serves in Step 2. Finally, by (47) and $\Lambda_n^2 = \Gamma_{n,\delta}/(n\gamma)$, together with $\Lambda_n \leq \sqrt{c} \leq 1$ from (40) and (39),

$$8\sqrt{\frac{\|K\|_\infty C_S \Xi}{n\gamma}} \leq 8\sqrt{\|K\|_\infty C_S C_0}\,\Lambda_n, \qquad \frac{16\|K\|_\infty \Xi}{3n\gamma} \leq \frac{16\|K\|_\infty C_0}{3}\Lambda_n^2 \leq \frac{16\|K\|_\infty C_0}{3}\Lambda_n.$$

Both requirements in (46) are therefore met by $u = C_*\Lambda_n$ as soon as

$$C_* \;\geq\; 8\sqrt{\|K\|_\infty C_S C_0} \qquad \text{and} \qquad C_* \;\geq\; \frac{16\|K\|_\infty C_0}{3}, \tag{48}$$

and under (48) the first bound in (37) fails with probability at most $\delta/3$.

Step 2: the denominator and the marginal empirical mean. For the second bound in (37), fix $s \in \mathcal{S}_n$ and choose j with $|s - s_j| \leq \zeta_{\mathcal{S}}$. By (42), applied under $\mathbb{E}_n$ and under $\mathbb{E}$,

$$|(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s)| \leq |(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s_j)| + \frac{u}{8},$$

so

$$\mathbb{P}\left(\sup_{s\in\mathcal{S}_n} |(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s)| > u\right) \leq \sum_{j=1}^{N_2} \mathbb{P}\left(|(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s_j)| > \tfrac{u}{2}\right).$$

The variable $W' := K_\gamma(S,s_j)$ satisfies $0 \leq W' \leq \|K\|_\infty/\gamma \leq U$ and $\mathrm{Var}(W') \leq \mathbb{E}W'^2 \leq (\|K\|_\infty/\gamma)\, q_\gamma(s_j) \leq \|K\|_\infty C_S/\gamma \leq \sigma^2$. Equation (27) with $u/2$ in place of u therefore bounds each summand by $2\exp\{-nu^2/(8\sigma^2 + 4Uu/3)\}$, exactly as in Step 1, and the right-hand side is at most $2N_2\exp(-\Xi) \leq \delta/3$ by (45) and $N_1 \geq 1$. Hence, under (48), the second bound in (37) fails with probability at most $\delta/3$.

For the third bound, let $g_1', \ldots, g_{N_3}'$ be a $(u/4)$-net of $\mathcal{G}_n$ in $\|\cdot\|_\infty$ with $0 \leq g_k' \leq 2$. Since $u/4 \leq 1$, Lemma 3(ii) gives $N_3 \leq (36A_n/u)^{\mathfrak{v}_n + d + 1}$. For $g \in \mathcal{G}_n$ choose k with $\|g - g_k'\|_\infty \leq u/4$, so that $|(\mathbb{E}_n - \mathbb{E})g| \leq |(\mathbb{E}_n - \mathbb{E})g_k'| + u/2$, and hence

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}_n} |(\mathbb{E}_n - \mathbb{E})g| > u\right) \leq \sum_{k=1}^{N_3} \mathbb{P}\left(|(\mathbb{E}_n - \mathbb{E})g_k'| > \tfrac{u}{2}\right).$$

The variable $W'' := g_k'(\mathbf{X})$ satisfies $0 \leq W'' \leq 2$ and $\mathrm{Var}(W'') \leq 1$, so (27) with $u/2$ in place of u, $U = 2$ and $\sigma^2 = 1$ bounds each summand by $2\exp\{-nu^2/(8 + 8u/3)\}$. As in Step 1, the right-hand side is at most $\delta/3$ provided that $nu^2 \geq 16\Xi_3$ and $nu^2 \geq 16\Xi_3 u/3$, where $\Xi_3 := \log(6N_3/\delta)$, that is provided that

$$u \;\geq\; 4\sqrt{\frac{\Xi_3}{n}} \qquad \text{and} \qquad u \;\geq\; \frac{16\Xi_3}{3n}.$$

Since $u \geq n^{-1}$ and $\log(A_n n/\gamma) \geq 1$, we have $\log N_3 \leq (\mathfrak{v}_n + d + 1)\log(36A_n n) \leq 2(\mathfrak{v}_n + d)(1 + \log 36)\log(A_n n/\gamma)$, so $\Xi_3 \leq C_0\Gamma_{n,\delta}$ with the constant C_0 of (47). Using $\gamma \leq 1$ and $\Lambda_n \leq 1$,

$$4\sqrt{\frac{\Xi_3}{n}} \leq 4\sqrt{\frac{C_0\Gamma_{n,\delta}}{n\gamma}} = 4\sqrt{C_0}\,\Lambda_n, \qquad \frac{16\Xi_3}{3n} \leq \frac{16C_0\Gamma_{n,\delta}}{3n\gamma} = \frac{16C_0}{3}\Lambda_n^2 \leq \frac{16C_0}{3}\Lambda_n,$$

so both conditions hold for $u = C_*\Lambda_n$ as soon as $C_* \geq 4\sqrt{C_0}$ and $C_* \geq 16C_0/3$. We now fix

$$C_* := \max\left\{8\sqrt{\|K\|_\infty C_S C_0},\; \frac{16\|K\|_\infty C_0}{3},\; 4\sqrt{C_0},\; \frac{16C_0}{3}\right\}, \tag{49}$$

which is at least $16/3 > 1$ because $C_0 \geq 1$, and which depends only on c_S, C_S and K. With this choice each of the three bounds in (37) fails with probability at most $\delta/3$, and a union bound over the three events yields $\mathbb{P}(\Omega_n) \geq 1 - \delta$.

Step 3: the denominator is bounded away from zero. On Ω_n, for $s \in \mathcal{S}_n$, the second bound in (37), the lower bound $q_\gamma(s) \geq 3c_S/4$ of Lemma 5, the bound $\Lambda_n \leq \sqrt{c}$ of (40) and $C_*\sqrt{c} \leq c_S/4$ from (39) give

$$\mathbb{E}_n K_\gamma(S,s) \geq q_\gamma(s) - |(\mathbb{E}_n - \mathbb{E})K_\gamma(S,s)| \geq q_\gamma(s) - C_*\Lambda_n \geq \frac{3c_S}{4} - C_*\sqrt{c} \geq \frac{3c_S}{4} - \frac{c_S}{4} = \frac{c_S}{2},$$

which is (38). □

A.4 Proof of Theorem 1

Proof. Let C_* and c be the constants of Lemma 6 and let Ω_n be the event in (37), so that $\mathbb{P}(\Omega_n) \geq 1-\delta$ whenever $\Gamma_{n,\delta} \leq cn\gamma$. We show that (17) holds on Ω_n.

On Ω_n fix $g \in \mathcal{G}_n$ and $s \in \mathcal{S}_n$, and write $\widehat{b}_{g,\gamma}(s) := \mathbb{E}_n\{K_\gamma(S,s)g(\mathbf{X})\}$ and $\widehat{q}_\gamma(s) := \mathbb{E}_n K_\gamma(S,s)$ for the numerator and the denominator of $\widehat{m}_{g,\gamma}(s)$ in (25), so that

$$\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s) = (\mathbb{E}_n - \mathbb{E})\{K_\gamma(S,s)g(\mathbf{X})\}, \qquad \widehat{q}_\gamma(s) - q_\gamma(s) = (\mathbb{E}_n - \mathbb{E})K_\gamma(S,s).$$

By (38), $\widehat{q}_\gamma(s) \geq c_S/2$, so $\widehat{m}_{g,\gamma}(s)$ is well defined. Using $\widehat{m}_{g,\gamma} = \widehat{b}_{g,\gamma}/\widehat{q}_\gamma$ and $m_{g,\gamma} = b_{g,\gamma}/q_\gamma$, we obtain

$$\begin{aligned}
\widehat{m}_{g,\gamma}(s) - m_{g,\gamma}(s) &= \frac{q_\gamma(s)\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)\widehat{q}_\gamma(s)}{q_\gamma(s)\widehat{q}_\gamma(s)} \\
&= \frac{q_\gamma(s)\{\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)\} - b_{g,\gamma}(s)\{\widehat{q}_\gamma(s) - q_\gamma(s)\}}{q_\gamma(s)\widehat{q}_\gamma(s)} \\
&= \frac{\{\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)\} - m_{g,\gamma}(s)\{\widehat{q}_\gamma(s) - q_\gamma(s)\}}{\widehat{q}_\gamma(s)}.
\end{aligned} \tag{50}$$

Since $0 \leq g \leq 2$ and $K_\gamma \geq 0$, we have $0 \leq b_{g,\gamma}(s) \leq 2q_\gamma(s)$ and hence $0 \leq m_{g,\gamma}(s) \leq 2$. On Ω_n the first two bounds in (37) give $|\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)| \leq C_*\Lambda_n$ and $|\widehat{q}_\gamma(s) - q_\gamma(s)| \leq C_*\Lambda_n$, so (50) and $\widehat{q}_\gamma(s) \geq c_S/2$ yield

$$\begin{aligned}
|\widehat{m}_{g,\gamma}(s) - m_{g,\gamma}(s)| &\leq \frac{|\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)| + m_{g,\gamma}(s)\,|\widehat{q}_\gamma(s) - q_\gamma(s)|}{\widehat{q}_\gamma(s)} \\
&\leq \frac{2}{c_S}\left(|\widehat{b}_{g,\gamma}(s) - b_{g,\gamma}(s)| + 2|\widehat{q}_\gamma(s) - q_\gamma(s)|\right) \\
&\leq \frac{2}{c_S}\left(C_*\Lambda_n + 2C_*\Lambda_n\right) = \frac{6C_*}{c_S}\Lambda_n.
\end{aligned}$$

Combining the preceding bound with (31) and the triangle inequality yields

$$\begin{aligned}
\sup_{g\in\mathcal{G}_n}\sup_{s\in\mathcal{S}_n}|\widehat{m}_{g,\gamma}(s) - m_g(s)| &\leq \sup_{g,s}|\widehat{m}_{g,\gamma}(s) - m_{g,\gamma}(s)| + \sup_{g,s}|m_{g,\gamma}(s) - m_g(s)| \\
&\leq \frac{6C_*}{c_S}\Lambda_n + C\gamma^2 \leq C\, r_{n,\delta}.
\end{aligned} \tag{51}$$

Now fix $h \in \mathcal{E}_n$. Apply (28) with index set $\mathcal{T} = \mathcal{V}_{\mathrm{R}} \times \mathcal{S}_n$ and, for $t = (v,s)$ and $g = v \circ h \in \mathcal{G}_n$,

$$F_1(t) := \widehat{m}_{g,\gamma}(s) - \mathbb{E}_n g, \qquad F_2(t) := m_g(s) - \mathbb{E}g,$$

both bounded by 2 in absolute value. By (26), $\sup_{\mathcal{T}}|F_1| = \widehat{D}_n(h)$ and $\sup_{\mathcal{T}}|F_2| = D_n(h)$. Since $F_1(t) - F_2(t) = \{\widehat{m}_{g,\gamma}(s) - m_g(s)\} - (\mathbb{E}_n - \mathbb{E})g$,

$$\begin{aligned}
\left|\widehat{D}_n(h) - D_n(h)\right| &\leq \sup_{t\in\mathcal{T}}|F_1(t) - F_2(t)| \\
&\leq \sup_{g\in\mathcal{G}_n}\sup_{s\in\mathcal{S}_n}|\widehat{m}_{g,\gamma}(s) - m_g(s)| + \sup_{g\in\mathcal{G}_n}|(\mathbb{E}_n - \mathbb{E})g| \\
&\leq C r_{n,\delta} + C_*\Lambda_n \leq C r_{n,\delta},
\end{aligned}$$

where the third inequality uses (51) and the third bound in (37), and the last uses $\Lambda_n \leq r_{n,\delta}$. The resulting bound does not depend on h, so it also bounds $\sup_{h\in\mathcal{E}_n}\left|\widehat{D}_n(h) - D_n(h)\right|$. This proves (17) on Ω_n, and hence with probability at least $1-\delta$.

For the rate, suppose that A_n and $\mathfrak{v}_n$ are bounded, $\delta_n = n^{-c_1}$, $\gamma_n \asymp (\log n/n)^{1/5}$ and $\epsilon_n \asymp 1/\log n$. Then $\log(A_n n/\gamma_n) = \log A_n + \log n + \frac{1}{5}\log(n/\log n) + O(1) \asymp \log n$ and $\log(1/\delta_n) = c_1 \log n$, so $\Gamma_{n,\delta_n} \asymp \log n$. With this bandwidth,

$$\gamma_n^2 \asymp \left(\frac{\log n}{n}\right)^{2/5}, \qquad \sqrt{\frac{\Gamma_{n,\delta_n}}{n\gamma_n}} \asymp \sqrt{\frac{\log n}{n^{4/5}(\log n)^{1/5}}} = \left(\frac{\log n}{n}\right)^{2/5},$$

so $r_{n,\delta_n} = O\{(\log n/n)^{2/5}\}$. The condition $\Gamma_{n,\delta_n} \leq cn\gamma_n$ reads $\log n = O\{n^{4/5}(\log n)^{1/5}\}$, and $\gamma_n \leq \epsilon_n$ reads $(\log n/n)^{1/5} = O(1/\log n)$. Both hold for all large n. □

A.5 Proof of Corollary 2

Proof. Let Ω_n be the event of Lemma 6, on which (17) holds for every $h \in \mathcal{E}_n$ simultaneously, so that

$$D_n(h) \le \widehat{D}_n(h) + Cr_{n,\delta}, \qquad h \in \mathcal{E}_n. \tag{52}$$

Fix $h \in \mathcal{E}_n$, $B > 0$, $\beta > 0$ with $\beta \neq (d+3)/2$, and $s \in \mathcal{S}_n$. Write $\mathbb{P}_s := \mathbb{P}_{h(\mathbf{X})|S=s}$ and $\mathbb{P}_0 := \mathbb{P}_{h(\mathbf{X})}$, which are probability measures on $\mathbb{B}^d$ because h takes values in $\mathbb{B}^d$ by Assumption 4. Let $f \in \mathcal{H}_d^\beta(B)$. Then $B^{-1}f \in \mathcal{H}_d^\beta$, so (1) with $\mathcal{V} = \mathcal{H}_d^\beta$ gives $|\mathbb{E}[f(h(\mathbf{X})) \mid S=s] - \mathbb{E}[f(h(\mathbf{X}))]| = B\left|\int B^{-1}f \,\mathrm{d}\mathbb{P}_s - \int B^{-1}f \,\mathrm{d}\mathbb{P}_0\right| \le B\,\mathrm{IPM}_{\mathcal{H}_d^\beta}(\mathbb{P}_s, \mathbb{P}_0)$. Since $d \ge 3$ and $\beta \neq (d+3)/2$, Theorem 2 of Park et al. (2026), restated in (8), then gives

$$|\mathbb{E}[f(h(\mathbf{X})) \mid S=s] - \mathbb{E}[f(h(\mathbf{X}))]| \le B\,\mathrm{IPM}_{\mathcal{H}_d^\beta}(\mathbb{P}_s, \mathbb{P}_0) \le C_{d,\beta} B\,\mathrm{IPM}_{\mathcal{V}_\mathrm{R}}(\mathbb{P}_s, \mathbb{P}_0)^{\alpha_{d,\beta}}. \tag{53}$$

The map $t \mapsto t^{\alpha_{d,\beta}}$ is nondecreasing on $[0,\infty)$, and $\mathrm{IPM}_{\mathcal{V}_\mathrm{R}}(\mathbb{P}_s, \mathbb{P}_0) \le D_n(h)$ by the definition of $D_n(h)$. The right-hand side of (53) is therefore at most $C_{d,\beta}BD_n(h)^{\alpha_{d,\beta}}$, which depends on neither s nor f. Take the supremum over $s \in \mathcal{S}_n$ and then over $f \in \mathcal{H}_d^\beta(B)$ on the left-hand side of (53). By the definition of $\mathrm{GDP}_{\infty,n}$ in (15) the result is $\sup_f \mathrm{GDP}_{\infty,n}(f \circ h)$, and applying (52) together with the monotonicity of $t \mapsto t^{\alpha_{d,\beta}}$ gives

$$\sup_{f \in \mathcal{H}_d^\beta(B)} \mathrm{GDP}_{\infty,n}(f \circ h) \le C_{d,\beta} B\, D_n(h)^{\alpha_{d,\beta}} \le C_{d,\beta} B \left\{\widehat{D}_n(h) + Cr_{n,\delta}\right\}^{\alpha_{d,\beta}},$$

which is (18). The event Ω_n depends on neither h, nor B, nor β, so the bound holds simultaneously for all of them. □

B Details of Training supFR

This appendix supplies the details of the learning procedure in Section 3.4. Appendix B.1 describes the updates of the candidate sensitive values and the discriminator parameters together with the descent step for the encoder and the prediction head. Appendix B.2 derives an equivalent formulation of the objective $\mathcal{L}(\phi, \psi)$ of Section 3.4 for mini-batch training. The mini-batch gradients are unbiased when the inner maximization variables, namely the sensitive value, the discriminator parameters and a sign introduced in that appendix, are fixed before the mini-batch is drawn. Appendix B.3 presents the batch and mini-batch procedures in Algorithms 1 and 2, respectively.

B.1 Optimization Procedure

In practice the encoder h_ϕ is a neural network followed by the normalization $\mathbf{z} \mapsto \mathbf{z}/\sqrt{\|\mathbf{z}\|_2^2 + 1} \in \mathbb{B}^d$, so that $\mathbf{Z} = h_\phi(\mathbf{X})$ takes values in $\mathbb{B}^d$, on which the ReLU discriminator class $\mathcal{V}_\mathrm{R}$ of (6) is defined. Minimizing $\mathcal{L}(\phi, \psi)$ of Section 3.4 is a min–max problem, which we solve by alternating gradient descent in (ϕ, ψ) with compass search in s (Kolda et al., 2003) and projected gradient ascent in (θ, μ). To describe the search and the ascent steps, write $\mathbf{Z}_i := h_\phi(\mathbf{X}_i)$ and, for $s \in \mathcal{S}$ and $(\theta, \mu) \in \mathbb{S}^{d-1} \times [-1, 1]$, let

$$\Delta(s; \theta, \mu) := \left| \sum_{i=1}^n \left(\widehat{w}_\gamma(i; s) - \frac{1}{n} \right) (\theta^\top \mathbf{Z}_i + \mu)_+ \right|. \tag{54}$$

Then (14) and (13) read

$$\mathcal{L}(\phi, \psi; s, \theta, \mu) = \frac{1}{n} \sum_{i=1}^n \ell(f_\psi(\mathbf{Z}_i), Y_i) + \lambda \Delta(s; \theta, \mu),$$

$$\widehat{\mathrm{supIPM}}_{\mathcal{V}_\mathrm{R}}^\gamma(\mathbf{Z}; S) = \sup_{s \in \mathcal{S}} \sup_{\theta \in \mathbb{S}^{d-1},\, \mu \in [-1,1]} \Delta(s; \theta, \mu).$$

Because the compass search and the gradient ascent can converge to a local maximum, we follow Park et al. (2026) and run them from several initial values in parallel.

Let $\mathbf{s} := (s^{(1)}, \dots, s^{(T)})$ be T candidate sensitive values and $\mathbf{v} := ((\theta_1, \mu_1), \dots, (\theta_M, \mu_M))$ be M discriminators, where $(\theta_m, \mu_m) \in \mathbb{S}^{d-1} \times [-1, 1]$ for $m \in [M] := \{1, \dots, M\}$, and let $s^\star := \arg\max_{s \in \{s^{(1)}, \dots, s^{(T)}\}} \max_{m \in [M]} \Delta(s; \theta_m, \mu_m)$. The candidates are updated by compass search on Q and the discriminators by gradient ascent on H, where

$$Q(s) := \max_{m \in [M]} \Delta(s; \theta_m, \mu_m), \qquad H(\mathbf{v}) := \sum_{m=1}^{M} \Delta(s^\star; \theta_m, \mu_m), \tag{55}$$

and $\mathbf{v}$ and $s^\star$ are held fixed at their current values in Q and in H, respectively.

We write Π_A for the projection onto a set A. One compass search step with step size ι evaluates Q at the candidate $s^{(t)}$ and at the two points $\Pi_{\mathcal{S}}(s^{(t)} \pm \iota)$, moves the candidate to the point with the largest value, and halves ι. The T candidates are moved in parallel. The search starts from the step size ι_0 and stops once $\iota < \gamma/8$.

Since the m-th summand of H depends on $\mathbf{v}$ only through (θ_m, μ_m), one ascent step on H updates all M discriminator parameters, and after each step $\Pi_{\mathbb{S}^{d-1} \times [-1,1]}$ projects each θ_m onto the unit sphere and each μ_m to $[-1, 1]$.

The encoder and the prediction head are then updated by gradient descent on

$$\max_{m \in [M]} \mathcal{L}(\phi, \psi; s^\star, \theta_m, \mu_m),$$

that is, on $\mathcal{L}(\phi, \psi)$ with the supremum over s restricted to the T candidates and the supremum over (θ, μ) to the M discriminators.

B.2 Mini-batch Learning

The updates of Appendix B.1 use all n observations. For mini-batch training, the absolute value and the supremum in (13) prevent the usual argument that an average over a random mini-batch is unbiased for the average over the sample. We therefore express the absolute value using the identity $|x| = \max_{\chi \in \{-1,1\}} \chi x$. We then estimate the resulting objective on a mini-batch that is drawn with the encoder, the prediction head and the inner variables, that is, the sensitive value, the discriminator parameters and the sign, held fixed. These steps yield an equivalent stochastic min–max formulation of $\min_{\phi,\psi} \mathcal{L}(\phi, \psi)$.

Throughout this subsection, we condition on the sample $\mathcal{D}$, and the bandwidth γ is fixed. Write $[n] := \{1, \dots, n\}$, let the mini-batch $\mathcal{B}$ be drawn uniformly from the subsets of $[n]$ of a fixed size $|\mathcal{B}|$, and let $\mathbb{E}_{\mathcal{B}}$ denote the expectation over this draw. As in the proof of Theorem 1, write $\widehat{q}_\gamma(s) := n^{-1} \sum_{j=1}^{n} K_\gamma(S_j, s)$, so that $\widehat{w}_\gamma(i; s) = K_\gamma(S_i, s)/\{n\widehat{q}_\gamma(s)\}$ when $\widehat{q}_\gamma(s) > 0$, and we set $\widehat{w}_\gamma(i; s) := 0$ when $\widehat{q}_\gamma(s) = 0$. The normalization $\widehat{q}_\gamma(s)$ uses all n sensitive values and does not depend on the encoder, so at a given s it costs $O(n)$ scalar kernel evaluations and no forward pass through h_ϕ.

An equivalent objective. Let $\chi \in \{-1, 1\}$ be a sign, and collect the inner variables in $\boldsymbol{\vartheta} := (s, \theta, \mu, \chi) \in \Theta := \mathcal{S} \times \mathbb{S}^{d-1} \times [-1, 1] \times \{-1, 1\}$. For a mini-batch $\mathcal{B} \subseteq [n]$, with $\mathbf{Z}_i = h_\phi(\mathbf{X}_i)$ and the discriminator $v_{\theta,\mu}$ of (6), define

$$\begin{aligned} \Delta_{\mathcal{B}}(s; \theta, \mu) &:= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \big(n\widehat{w}_\gamma(i; s) - 1\big)\, v_{\theta,\mu}(\mathbf{Z}_i), \\ \mathcal{J}_{\mathcal{B}}(\phi, \psi; \boldsymbol{\vartheta}) &:= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \ell(f_\psi(\mathbf{Z}_i), Y_i) + \lambda\chi\, \Delta_{\mathcal{B}}(s; \theta, \mu), \end{aligned} \tag{56}$$

and write $\mathcal{J} := \mathcal{J}_{[n]}$ for the batch objective. The factor n denotes the full sample size, and the weights $\widehat{w}_\gamma(i; s)$ are computed using all n sensitive values. Since $\Delta_{[n]}(s; \theta, \mu)$ is the signed sum inside the absolute value in (54), we have $\Delta(s; \theta, \mu) = |\Delta_{[n]}(s; \theta, \mu)|$, and the identity $|x| = \max_{\chi \in \{-1,1\}} \chi x$ together with $\lambda \geq 0$ gives

$$\mathcal{L}(\phi, \psi) = \sup_{\boldsymbol{\vartheta} \in \Theta} \mathcal{J}(\phi, \psi; \boldsymbol{\vartheta}). \tag{57}$$

With h_ϕ fixed, the representation of an input must not depend on the mini-batch. Therefore, the normalization of the encoder's output cannot use mini-batch statistics. The pointwise normalization of Appendix B.1 satisfies this requirement.

Proposition 7 (Unbiased mini-batch objective and gradients)**.** *For every fixed* $(\phi, \psi, \boldsymbol{\vartheta})$*,*

$$\mathbb{E}_{\mathcal{B}}\big[\mathcal{J}_{\mathcal{B}}(\phi, \psi; \boldsymbol{\vartheta})\big] = \mathcal{J}(\phi, \psi; \boldsymbol{\vartheta}), \qquad \mathcal{L}(\phi, \psi) = \sup_{\boldsymbol{\vartheta} \in \Theta} \mathbb{E}_{\mathcal{B}}\big[\mathcal{J}_{\mathcal{B}}(\phi, \psi; \boldsymbol{\vartheta})\big]. \tag{58}$$

If $\ell(f_\psi(\mathbf{Z}_i), Y_i)$ *and* $v_{\theta,\mu}(\mathbf{Z}_i)$ *are differentiable in one of* ϕ*,* ψ*,* θ *and* μ *for every* $i \in [n]$*, with the other variables held fixed, then the gradient* ∇ *with respect to that variable satisfies*

$$\mathbb{E}_{\mathcal{B}}\big[\nabla \mathcal{J}_{\mathcal{B}}(\phi, \psi; \boldsymbol{\vartheta})\big] = \nabla \mathcal{J}(\phi, \psi; \boldsymbol{\vartheta}). \tag{59}$$

Suppose that $\mathcal{B}$ *is drawn uniformly and independently of the random parameters* $(\phi, \psi, \boldsymbol{\vartheta})$*. Then both statements hold conditional on* $(\phi, \psi, \boldsymbol{\vartheta})$*. Finally, suppose that* $\boldsymbol{\vartheta}^\star$ *maximizes* $\mathcal{J}(\phi, \psi; \cdot)$ *over* Θ *and is fixed before* $\mathcal{B}$ *is drawn, that* $\mathcal{L}$ *is differentiable at* (ϕ, ψ)*, and that the condition above holds for* ϕ *and* ψ *at* $\boldsymbol{\vartheta}^\star$*. Then*

$$\mathbb{E}_{\mathcal{B}}\big[\nabla_{(\phi,\psi)} \mathcal{J}_{\mathcal{B}}(\phi, \psi; \boldsymbol{\vartheta}^\star)\big] = \nabla_{(\phi,\psi)} \mathcal{L}(\phi, \psi). \tag{60}$$

Proof. Fix $(\phi, \psi, \boldsymbol{\vartheta})$. By (56), $\mathcal{J}_{\{i\}} = \ell(f_\psi(\mathbf{Z}_i), Y_i) + \lambda\chi\big(n\widehat{w}_\gamma(i; s) - 1\big)v_{\theta,\mu}(\mathbf{Z}_i)$ for the single observation i, so that $\mathcal{J}_{\mathcal{B}} = |\mathcal{B}|^{-1}\sum_{i\in\mathcal{B}} \mathcal{J}_{\{i\}}$ and $\mathcal{J} = n^{-1}\sum_{i=1}^n \mathcal{J}_{\{i\}}$. Under uniform sampling, $\mathbb{P}(i \in \mathcal{B}) = |\mathcal{B}|/n$ for every $i \in [n]$, so

$$\mathbb{E}_{\mathcal{B}}\mathcal{J}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^n \mathbb{P}(i \in \mathcal{B})\, \mathcal{J}_{\{i\}} = \frac{1}{n}\sum_{i=1}^n \mathcal{J}_{\{i\}} = \mathcal{J}.$$

The second identity in (58) then follows from (57). The expectation $\mathbb{E}_{\mathcal{B}}$ is a finite average over the subsets of $[n]$ of size $|\mathcal{B}|$, so it commutes with differentiation, and the same calculation with $\nabla\mathcal{J}_{\{i\}}$ in place of $\mathcal{J}_{\{i\}}$ gives (59). If $\mathcal{B}$ is independent of a random $(\phi, \psi, \boldsymbol{\vartheta})$, conditioning on $(\phi, \psi, \boldsymbol{\vartheta})$ leaves $\mathbb{P}(i \in \mathcal{B}) = |\mathcal{B}|/n$, which gives the conditional statements. For the last assertion, the map $(\phi', \psi') \mapsto \mathcal{L}(\phi', \psi') - \mathcal{J}(\phi', \psi'; \boldsymbol{\vartheta}^\star)$ is nonnegative by (57) and vanishes at (ϕ, ψ), where it therefore attains its minimum. Its gradients with respect to ϕ and to ψ at (ϕ, ψ) therefore vanish, so $\nabla_{(\phi,\psi)}\mathcal{L}(\phi, \psi) = \nabla_{(\phi,\psi)}\mathcal{J}(\phi, \psi; \boldsymbol{\vartheta}^\star)$, and (59) completes the proof. □

We fix the sensitive value, the discriminator and the sign before drawing the mini-batch used for the gradient step, so that the mini-batch gradient is unbiased for the gradient of $\mathcal{J}$ by Proposition 7, and Algorithm 2 follows this order. The MMD penalty of FREM biased on mini-batches, so convergence to a stationary point of the batch objective is not guaranteed (Lee et al., 2026), and it also requires a kernel evaluation for every pair of representations.

B.3 Algorithm

Algorithm 1 repeats the three updates of Appendix B.1 for R iterations. Each iteration performs J_v discriminator updates, each preceded by a compass search from newly initialized candidates, recomputes $s^\star$ with the updated discriminators, and concludes with one descent step. The compass search starts from the step size ι_0 and stops once $\iota < \gamma/8$, and the discriminator updates and the descent step use the step sizes α_v and α.

Algorithm 2 is the mini-batch version of Algorithm 1 and follows the order of sampling required by Proposition 7. The compass search, the candidate $s^\star$ and the signs $\chi_m := \operatorname{sign}\Delta_{\mathcal{B}}(s^\star; \theta_m, \mu_m)$, with $\operatorname{sign}(0) := 1$, are computed on one mini-batch with Δ replaced by $|\Delta_{\mathcal{B}}|$, and each gradient step uses a new mini-batch drawn independently of the preceding ones. The ascent step uses

$$H_{\mathcal{B}}(\mathbf{v}) := \sum_{m=1}^M \chi_m\, \Delta_{\mathcal{B}}(s^\star; \theta_m, \mu_m), \tag{61}$$

whose gradient is unbiased, by the argument of Proposition 7, for the gradient of

$$\sum_{m=1}^M \chi_m \Delta_{[n]}(s^\star; \theta_m, \mu_m).$$

When each χ_m equals the sign of $\Delta_{[n]}(s^\star;\theta_m,\mu_m)$, the sum in the preceding display equals $H(\mathbf{v})$ in (55), and if the M discrepancies $\Delta_{[n]}(s^\star;\theta_m,\mu_m)$ are all nonzero, the gradient of this sum equals $\nabla_{\mathbf{v}}H(\mathbf{v})$. The descent step uses $\mathcal{J}_{\mathcal{B}}$ of (56) at $(s^\star,\theta_{m^\star},\mu_{m^\star},\chi_{m^\star})$, where $m^\star$ indexes the discriminator with the largest $|\Delta_{\mathcal{B}}(s^\star;\theta_m,\mu_m)|$ on the preceding mini-batch.

Algorithm 1 supFR

Input: sample $\mathcal{D}$, supervised loss ℓ, kernel K_γ with bandwidth γ, penalty parameter λ, number of candidates T, number of discriminators M, number of iterations R, number of inner updates J_v, initial step size ι_0 of the compass search, step sizes α_v and α
Output: encoder h_ϕ and prediction head f_ψ

Initialize ϕ, ψ, and $\mathbf{v}$ with $(\theta_m,\mu_m)\in\mathbb{S}^{d-1}\times[-1,1]$ for $m\in[M]$
for $r=1,\dots,R$ **do**
  for $j=1,\dots,J_v$ **do**
    Initialize $\mathbf{s}$ at T distinct points of $\mathcal{S}$ and set $\iota\leftarrow\iota_0$
    while $\iota\ge\gamma/8$ **do**
      $s^{(t)}\leftarrow\arg\max_{s\in\{s^{(t)},\,\Pi_{\mathcal{S}}(s^{(t)}-\iota),\,\Pi_{\mathcal{S}}(s^{(t)}+\iota)\}}Q(s)$ for $t\in[T]$, and $\iota\leftarrow\iota/2$
    end while
    $s^\star\leftarrow\arg\max_{s\in\{s^{(1)},\dots,s^{(T)}\}}\max_{m\in[M]}\Delta(s;\theta_m,\mu_m)$
    $\mathbf{v}\leftarrow\Pi_{\mathbb{S}^{d-1}\times[-1,1]}(\mathbf{v}+\alpha_v\,\nabla_{\mathbf{v}}H(\mathbf{v}))$
  end for
  Recompute $s^\star$ by repeating lines 4–8 with the updated $\mathbf{v}$
  $(\phi,\psi)\leftarrow(\phi,\psi)-\alpha\,\nabla_{(\phi,\psi)}\big[\frac{1}{n}\sum_{i=1}^n\ell(f_\psi(\mathbf{Z}_i),Y_i)+\lambda\max_{m\in[M]}\Delta(s^\star;\theta_m,\mu_m)\big]$
end for
return h_ϕ, f_ψ

Algorithm 2 supFR with mini-batches

Input: the inputs of Algorithm 1 and a mini-batch size $|\mathcal{B}|$
Output: encoder h_ϕ and prediction head f_ψ

Initialize ϕ, ψ, and $\mathbf{v}$ with $(\theta_m,\mu_m)\in\mathbb{S}^{d-1}\times[-1,1]$ for $m\in[M]$
for $r=1,\dots,R$ **do**
  for $j=1,\dots,J_v$ **do**
    Draw a mini-batch $\mathcal{B}$ and run lines 4–8 of Algorithm 1 with Δ replaced by $|\Delta_{\mathcal{B}}|$
    $\chi_m\leftarrow\operatorname{sign}\Delta_{\mathcal{B}}(s^\star;\theta_m,\mu_m)$ for $m\in[M]$
    Draw a new mini-batch $\mathcal{B}$ and set $\mathbf{v}\leftarrow\Pi_{\mathbb{S}^{d-1}\times[-1,1]}(\mathbf{v}+\alpha_v\,\nabla_{\mathbf{v}}H_{\mathcal{B}}(\mathbf{v}))$
  end for
  Repeat lines 4–5 with the updated $\mathbf{v}$, and set $m^\star\leftarrow\arg\max_{m\in[M]}|\Delta_{\mathcal{B}}(s^\star;\theta_m,\mu_m)|$
  Draw a new mini-batch $\mathcal{B}$ and set $(\phi,\psi)\leftarrow(\phi,\psi)-\alpha\,\nabla_{(\phi,\psi)}\mathcal{J}_{\mathcal{B}}\big(\phi,\psi;s^\star,\theta_{m^\star},\mu_{m^\star},\chi_{m^\star}\big)$
end for
return h_ϕ, f_ψ

C Experiments on Synthetic Data

This appendix gives the data-generating process, the training settings and the detailed result of the simulation in Section 5.1.

C.1 Data-generating process

The data are generated as

$$\begin{aligned}
&\omega(s):=\exp\left\{-\frac{(s-s_0)^2}{2\tau^2}\right\}, && s_0=0.8,\quad \tau=0.043,\\
&S\sim p_S(\cdot), && p_S(s)\propto 1-0.9\,\omega(s),\quad s\in[0,1],\\
&\mathbf{X}_{\text{task}}\sim N(\mathbf{0},\mathbf{I}_4), && \mathbf{X}_{\text{noise}}\sim N(\mathbf{0},\mathbf{I}_3),\\
&X_{\text{loc}}:=3\,\omega(S)+0.5\,\varepsilon_1. &&
\end{aligned}\tag{62}$$

where $\varepsilon_1, \varepsilon_2 \sim N(0,1)$, and S, $\mathbf{X}_{\text{task}}$, $\mathbf{X}_{\text{noise}}$, ε_1 and ε_2 are mutually independent. The function ω is a Gaussian kernel centered at $s_0 = 0.8$. Since the density of S is smallest around s_0, the interval $[s_0 - \tau,\ s_0 + \tau]$ has probability approximately 0.022, so the dependence of X_{loc} on S introduced through ω is concentrated in a small part of the sensitive domain. Let $\boldsymbol{\upsilon} := (0.8, -0.6, 0.4, 0.5)^\top$ and let σ be the logistic function of Section 5. The outcome and the input are

$$\begin{aligned} Y &:= \sigma\big(\boldsymbol{\upsilon}^\top \mathbf{X}_{\text{task}} + \omega(S) + 0.3\,\varepsilon_2\big)\,, \\ \mathbf{X} &:= (\mathbf{X}_{\text{task}},\ \mathbf{X}_{\text{noise}}, X_{\text{loc}}) \in \mathbb{R}^8. \end{aligned} \tag{63}$$

The sensitive attribute enters the input only through X_{loc}. Since $Y \in (0,1)$, the task is regression with the squared-error loss, and prediction performance is measured by the test $1 - \text{MAE}$.

C.2 Settings

Table 2 gives the training settings of supFR in the notation of Algorithm 2. Following the experimental setting of Jiang et al. (2022), the bandwidth of the kernel K_γ is $\gamma = \widehat{\sigma}_S\,(3n/4)^{-1/5}$, where $\widehat{\sigma}_S$ is the sample standard deviation of $S_1, \ldots, S_n$. For each of five seeds, training and test samples of size $5{,}000$ and a validation sample of size $1{,}000$ are drawn. supFR is trained with the penalty parameter λ of (14) in $\{0, 0.01, 0.03, 0.1, 0.3, 1, 3\}$, and FREM with the penalty parameter λ of (5) in $\{0.05, 0.1, 0.2, 0.5, 1, 2\}$ at each of two kernel bandwidths and then selected using validation set. The representation-level disparity is computed as in Appendix D.4.

Table 2: Training settings for the simulation.

batch	epochs	learning rate	weight decay	M	J_v	α_v	T	ι_0
200	200	10^{-3}	0	100	20	10^{-2}	17	0.108

C.3 Result

In the left panel of Figure 2, each curve connects the models of one method trained with increasing values of λ. For each method, the circled point marks the model with the smallest representation-level disparity among those that retain the test $1 - \text{MAE}$ of the model trained without a fairness penalty, 0.952. The right panel shows $\widehat{\Delta}_{\hat{f}_1}(s)$ for the head $\hat{f}_1$ of (20) fitted on the representation of each circled model, averaged over the five seeds. The remaining disparity shows that a downstream head can still produce different mean predictions across sensitive values. The fluctuations away from s_0 are not reproduced across seeds, whereas the peak of FREM at s_0 is.

D Experiments on Real Datasets

This appendix supplements Section 5.2. Appendices D.1–D.4 describe the datasets, the training settings, the baselines and the evaluation measures. Appendix D.5 gives the results omitted from the main text, Appendix E compares the computational cost of supFR and FREM, and Appendix F examines the effect of the discriminator class.

D.1 Datasets

Table 3 summarizes the five datasets. The tabular datasets are Adult (Becker & Kohavi, 1996), Communities and Crime (Redmond, 2002) and ACSIncome, the income task of the 2018 American Community Survey constructed by Ding et al. (2021). On Adult and ACSIncome a large GDP_∞ indicates that the conditional mean prediction differs substantially from the marginal mean at some age, and on Crime that the predicted crime rate depends on the racial composition of a community. S is min–max scaled to $[0,1]$ on the training split. The methods for a binary S use S split at the median of the training split in Section 5.2, and Appendix D.5 also reports them with S split at its mean.

The graph datasets Pokec-z and Pokec-n are the social-network graphs drawn from the Pokec network of Takac & Zabovsky (2012), constructed with the loader of FairGNN (Dai & Wang, 2021)

Table 3: Datasets.

Dataset	Task	input dim.	n train / test	Y	S
Adult	classification	102	32,561 / 12,661	income $>$ \$50k	age
Crime	regression	122	1,794 / 200	violent crime rate	Black population share
ACSIncome	classification	9	125,226 / 39,133	income	age
Pokec-z	classification	276	33,898 / 16,949	spoken language	age
Pokec-n	classification	265	33,284 / 16,643	spoken language	age

and with the node counts reported by Kong et al. (2025). The encoder is a graph convolutional network (GCN) (Kipf & Welling, 2016) followed by a linear head. Age is missing for about 31% of the nodes, coded as zero. These nodes enter the supervised loss but not the fairness penalty and not the fairness measures, and S is scaled over the nodes with a known age. Every method is trained on the same node split and evaluated in the same manner, with the seed of the network initialization fixed in each run.

D.2 Training Settings

On the tabular datasets, for every method, the encoder is a two-layer network with 50 SELU units (Klambauer et al., 2017) per layer, followed by the normalization in Appendix B.1, so that $d = 50$, and the prediction head is linear. On the graph datasets, the encoder is the GCN of Appendix D.1. Table 4 gives the batch size, the number of epochs, the learning rate and the weight decay, which are common to all methods. For supFR, it gives the number M of discriminators and the number J_v of discriminator updates with step size α_v. The last two columns are the number T of candidate sensitive values and the initial step size ι_0 of the compass search over s, which stops once its step size falls below $\gamma/8$. The bandwidth γ follows the rule of Appendix C.2. The penalty parameter λ of (14) takes values between 0.003 and 3 for supFR. The running times and peak GPU memory usage in Appendix E are measured on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory under PyTorch 2.0.1 and CUDA 11.7.

Table 4: Training settings.

Dataset	batch	epochs	learning rate	weight decay	M	J_v	α_v	T	ι_0
Adult	1024	200	10^{-3}	10^{-2}	100	20	10^{-2}	17	0.108
Crime	200	200	10^{-3}	10^{-2}	100	20	10^{-2}	17	0.108
ACSIncome	1024	200	10^{-3}	10^{-2}	100	20	10^{-2}	17	0.108
Pokec-z / Pokec-n	1024	100	10^{-3}	0	100	10	10^{-2}	17	0.108

D.3 Baselines

Table 1 in Section 5.2 classifies the baselines by the sensitive attribute and the learning target of their original formulation. The baselines share the encoder, the prediction head and the training settings of Appendix D.2, and differ from supFR in their fairness term and in the range of its penalty parameter, as follows.

- FREM (Kong et al., 2025) penalizes the empirical EIPM of (4) with the RKHS unit ball of a Gaussian kernel of bandwidth 1.0. The kernel bandwidth on S is 0.03, 0.05 or 0.07 and choose one using validation set, and the penalty values lie between 0.01 and 3.
- Reg-GDP (Jiang et al., 2022) and Reg-HGR (Mary et al., 2019) penalize a kernel estimate of the GDP of (10) and an estimate of the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation between the prediction of the trained head and S, respectively, with penalty values between 0.02 and 100.
- ADV (Zhang et al., 2018), in the adaptation of Kong et al. (2025), trains an adversary to predict the continuous S from the representation, with penalty values between 5 and 5×10^4.
- LAFTR (Madras et al., 2018), MMD (Deka & Sutherland, 2023) and sIPM-LFR (Kim et al., 2022) align the representations of the two groups of a binary S through an adversary, the MMD and an IPM with sigmoid discriminators, respectively. S is split at its median in

Section 5.2 and at its mean in the additional panels of Appendix D.5, and the penalty values lie between 0.1 and 10.

D.4 Evaluation Measures

Prediction and fairness measures. All measures are computed on the test data, indexed here by $i = 1, \dots, n$. Prediction performance is the accuracy for classification and $1 - \mathrm{MAE}$ for regression. Fairness is measured by the prediction-level disparity $\mathrm{GDP}_\infty(f_\psi \circ h_\phi)$ and by the representation-level disparity $\mathrm{GDP}_\infty(h_\phi; \mathcal{F}_c)$ of (19), with $c = 1$ except in Figures 8 and 9. Both are computed from the estimates

$$\widehat{\Delta}_f(s) := \Big| \sum_{i=1}^{n} \widehat{w}_{\tilde{\gamma}}(i; s)\, f(\mathbf{Z}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{Z}_i) \Big|, \qquad \widehat{\mathrm{GDP}}_\infty(f \circ h) := \max_{s \in \mathcal{Q}} \widehat{\Delta}_f(s), \tag{64}$$

where $\widehat{w}_{\tilde{\gamma}}(i; s)$ is the kernel weight $\widehat{w}_\gamma(i; s)$ of Section 3.3 with a Gaussian kernel and the bandwidth $\tilde{\gamma}$, computed using the same bandwidth rule as in Appendix C.2, but with the test data in place of the training data, so that the first sum is the Nadaraya–Watson estimate of $\mathbb{E}[f(\mathbf{Z}) \mid S = s]$, and $\mathcal{Q}$ is a grid of spacing 10^{-3} of S.

Representation-level disparity. The head $\hat{f}_c$ of (20) is computed with $\widehat{\mathrm{GDP}}_\infty$ of (64) formed from the training sample, by projected gradient descent over $(\mathbf{w}, b)$ from eight initial heads, with $\mathbf{w}$ projected onto the ball of radius c after every step. The reported value $\widehat{\mathrm{GDP}}_\infty(h; \mathcal{F}_c) = \widehat{\mathrm{GDP}}_\infty(\hat{f}_c \circ h)$ of Section 5 is computed on the test sample.

Additional fairness measures. Appendix D.5 evaluates the methods under five further measures. The kernel GDP and the binned GDP are two estimates of the GDP of (10). The kernel GDP averages $\widehat{\Delta}_{f_\psi}(s)$ against a kernel density estimate of S, as in Kong et al. (2025), and the binned GDP averages, over the distinct observed values of S, the absolute difference between the mean prediction at that value and the overall mean. The empirical supIPM of (13) is computed on the test representation. The last two measure the dependence between the prediction $\hat{Y} := f_\psi(h_\phi(\mathbf{X}))$ and S: the HGR maximal correlation (Mary et al., 2019) and the mutual information,

$$\begin{aligned} \mathrm{HGR}(\hat{Y}, S) &:= \sup_{g_1, g_2} \mathrm{Corr}\big(g_1(\hat{Y}), g_2(S)\big), \\ I(\hat{Y}; S) &:= \int p_{\hat{Y},S}(\hat{y}, s) \log \frac{p_{\hat{Y},S}(\hat{y}, s)}{p_{\hat{Y}}(\hat{y})\, p_S(s)}\, \mathrm{d}\hat{y}\, \mathrm{d}s, \end{aligned} \tag{65}$$

where the supremum is over non-constant functions with finite second moments and $p_{\hat{Y},S}$, $p_{\hat{Y}}$ and p_S are the joint and the marginal densities. Both vanish if and only if $\hat{Y}$ and S are independent. The HGR is estimated as in Mary et al. (2019) from a kernel density estimate of the joint distribution of $(\hat{Y}, S)$, and the mutual information with the nearest-neighbor estimator of Kraskov et al. (2004) and Ross (2014).

In Section 5.2 and in this appendix, each point on a curve reports the mean prediction performance and the mean fairness measure over five runs trained with the same value of λ, and the vertical and the horizontal bars give one standard deviation over the five runs of the prediction performance and of the fairness measure, respectively. The black point is the model trained without a fairness penalty, at $\lambda = 0$.

D.5 Additional Results

This subsection collects the results omitted from Section 5.2. In every figure on the tabular datasets, the rows are Adult, Crime and ACSIncome from top to bottom; the left panels show the methods that use the continuous S, and the middle and the right panels show the methods that require a binary S, with S split at its median as in Section 5.2 and at its mean, respectively.

Prediction-level disparity. Figure 6 compares the methods by the prediction-level disparity on the three tabular datasets, with the binary-S methods under both splits of S. supFR shows the comparable performance against other baselines.

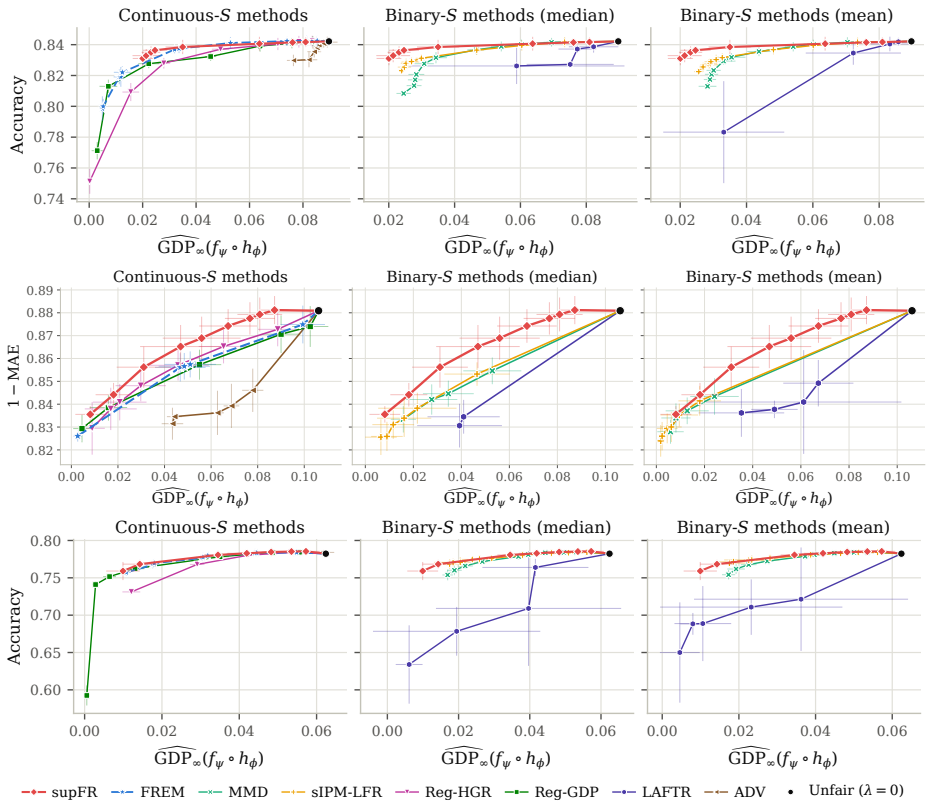


Figure 6: Prediction performance against the prediction-level disparity $\widehat{\mathrm{GDP}}_\infty(f_\psi \circ h_\phi)$.

Representation-level disparity. Figure 7 compares the methods by the representation-level disparity over $\mathcal{F}_1$ on the same datasets. At comparable prediction performance, supFR attains the lowest representation-level disparity on the three datasets. The supremum over s and over the discriminators constrains the representation at every sensitive value, whereas an average over S, or a penalty on the trained head, leaves a dependence on S that a downstream head can exploit.

Larger norm bounds. Figures 8 and 9 evaluate the representation-level disparity over $\mathcal{F}_4$ and $\mathcal{F}_8$, respectively; these larger norm bounds admit a broader class of downstream heads. At comparable prediction performance, supFR attains the lowest representation-level disparity on the three tabular datasets, and its margin over the baselines is larger than at $c = 1$, so its advantage persists across the evaluated norm bounds.

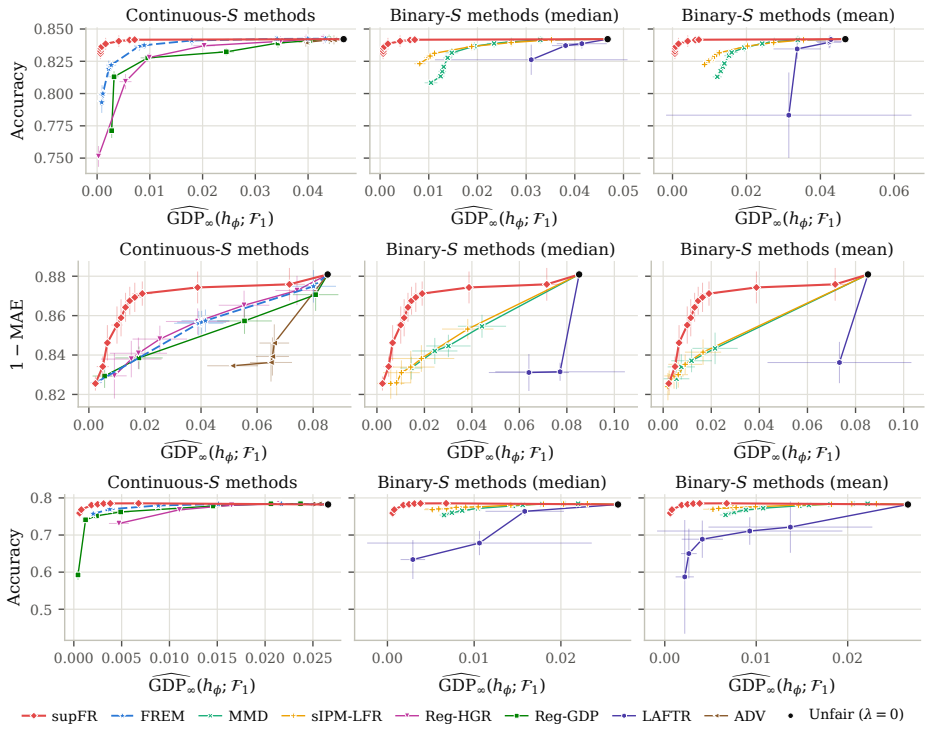


Figure 7: Prediction performance against the representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$.

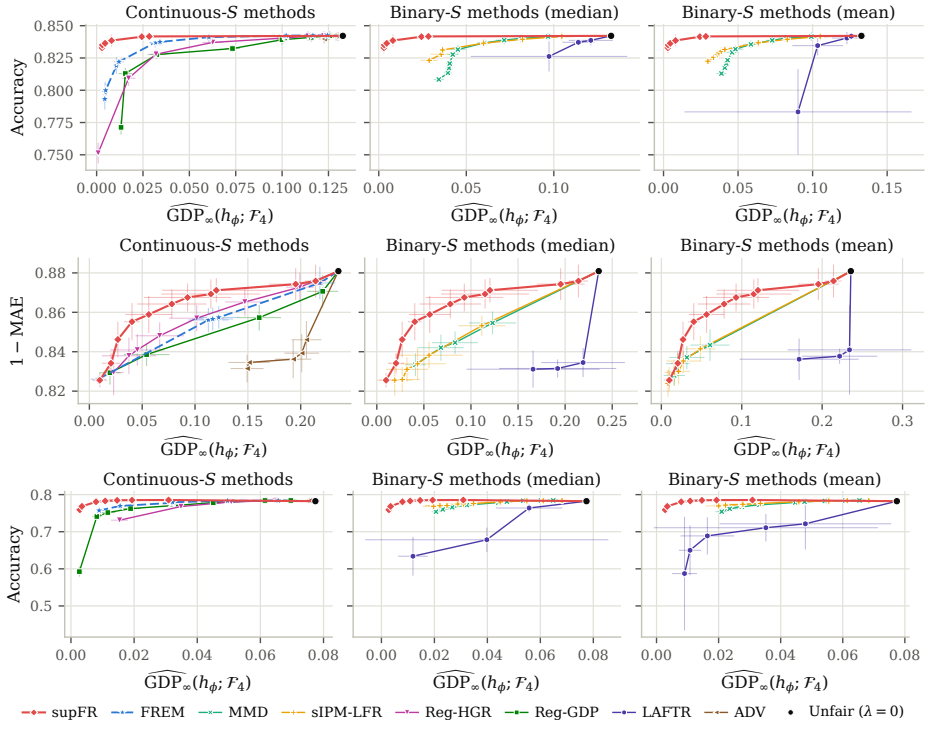


Figure 8: Prediction performance against the representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_4)$.

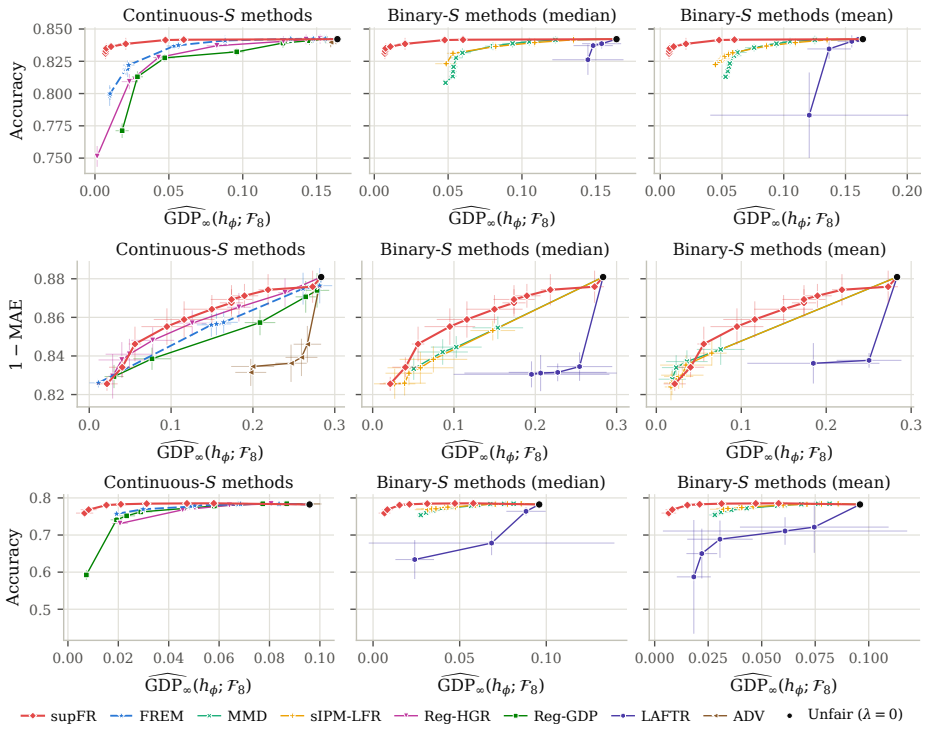


Figure 9: Prediction performance against the representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_8)$.

Matched prediction performance. Table 5 compares the methods at a fixed level of prediction performance on each tabular dataset. For each method and each disparity, the entry is the smallest disparity among the models of that method whose prediction performance lies within 0.004 of the level. A dash indicates that no model of that method meets this criterion, and the smallest value in each row is in bold. supFR attains the smallest value in every row. Under the representation-level disparity, its value is at least three times smaller than that of every baseline on the three datasets.

Table 5: Prediction-level disparity $\widehat{\mathrm{GDP}}_\infty(f_\psi \circ h_\phi)$ and representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$ on the tabular datasets at matched prediction performance.

	supFR	FREM	Reg-GDP	Reg-HGR	ADV	LAFTR	MMD	sIPM-LFR
Adult, accuracy 0.842								
prediction-level	**0.0350**	0.0529	0.0639	0.0706	0.0855	0.0822	0.0541	0.0601
representation-level ($c = 1$)	**0.0014**	0.0086	0.0339	0.0192	0.0388	0.0403	0.0186	0.0197
Crime, $1 - \mathrm{MAE}$ 0.855								
prediction-level	**0.0369**	0.0468	0.0550	0.0458	–	–	0.0529	0.0465
representation-level ($c = 1$)	**0.0100**	0.0338	0.0504	0.0315	–	–	0.0380	0.0314
ACSIncome, accuracy 0.780								
prediction-level	**0.0279**	0.0302	0.0350	0.0423	0.0617	–	0.0368	0.0281
representation-level ($c = 1$)	**0.0018**	0.0088	0.0146	0.0165	0.0248	–	0.0145	0.0099

Two-layer MLP heads. We also evaluate the two-layer heads

$$\mathcal{F}'_c := \big\{\sigma(\mathbf{w}_2^\top \,\mathrm{relu}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + b_2) : \ \|\mathbf{W}_1\|_2 \le 1,\ \|\mathbf{w}_2\|_2 \le c\big\},$$

which have the same Lipschitz bound as the heads of $\mathcal{F}_c$. Figure 10 reports the representation-level disparity over $\mathcal{F}'_1$. The heads have 16 hidden units and are fitted as in Appendix D.4, with $\mathbf{W}_1$ projected onto the spectral-norm ball and $\mathbf{w}_2$ onto the ball of radius c after every step. At comparable prediction performance, the values over $\mathcal{F}'_1$ are close to those over $\mathcal{F}_1$.

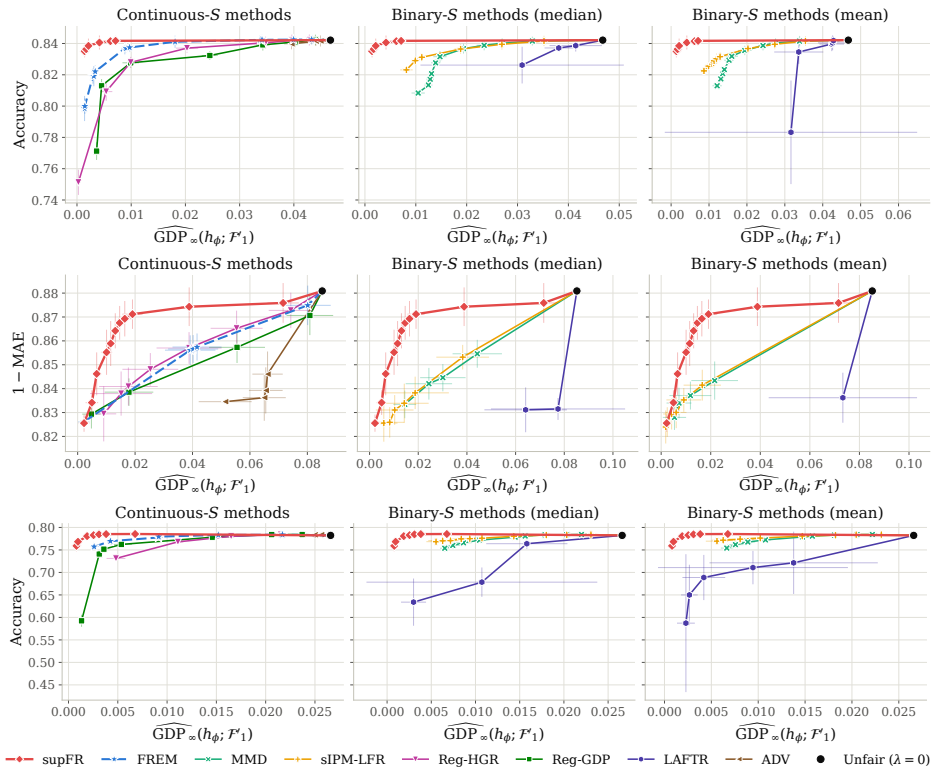


Figure 10: Prediction performance against the representation-level disparity over the two-layer heads, $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}'_1)$.

Alternative disparity measures. Figures 11–13 compare the methods by the kernel estimate of GDP, the binned estimate of GDP and the empirical supIPM, respectively, all defined in Appendix D.4. The two estimates of GDP evaluate the trained prediction head and average the disparity over S, whereas the empirical supIPM evaluates the representation on the test dataset. Under the two averages, supFR and FREM attain similar values. Under the empirical supIPM, supFR is markedly lower than every other method on Adult and Crime. This measure is the quantity that supFR penalizes, and Theorem 1 shows that it converges uniformly to the population supIPM of the representation.

Dependence of the fitted prediction on S. Figures 14 and 15 compare the methods by the HGR maximal correlation and by the mutual information between the prediction of the trained head and S. No method attains the smallest value on all three datasets, and supFR is comparable to FREM under both measures. These two measures depend on the trained head alone, so a method that penalizes that head can attain a small value while the representation still depends on S.

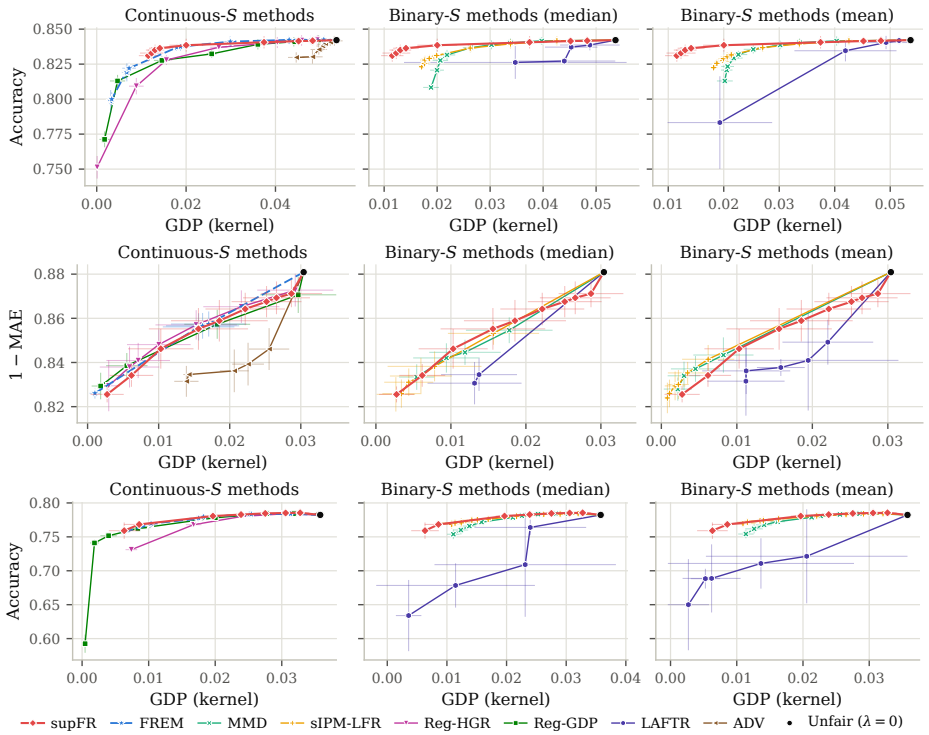


Figure 11: Prediction performance against the kernel estimate of GDP of the fitted prediction model.

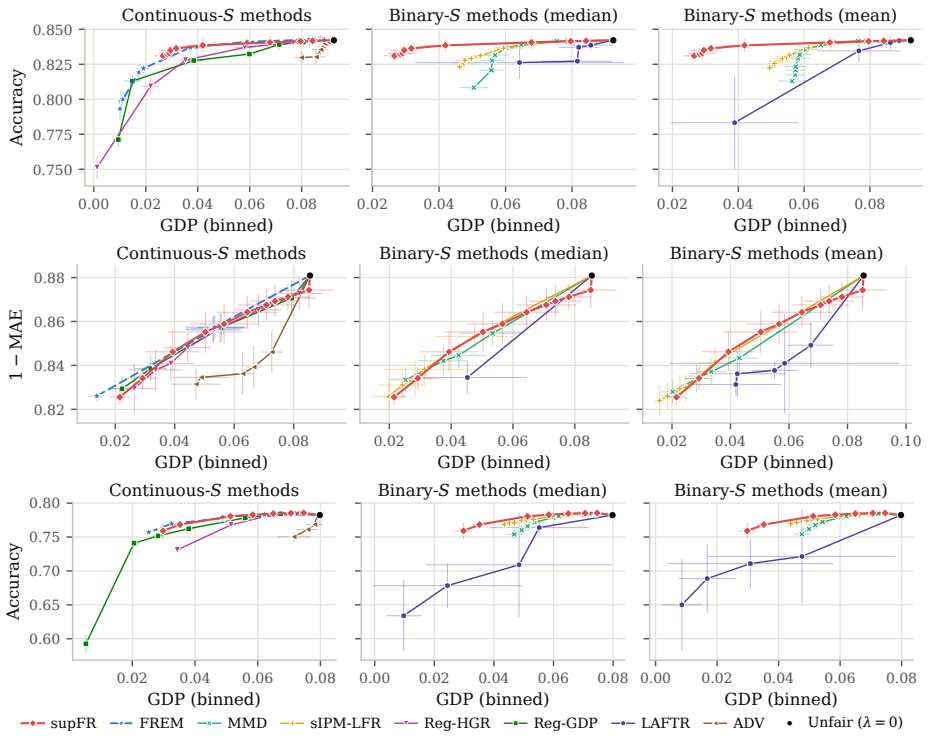


Figure 12: Prediction performance against the binned estimate of GDP of the fitted prediction model.

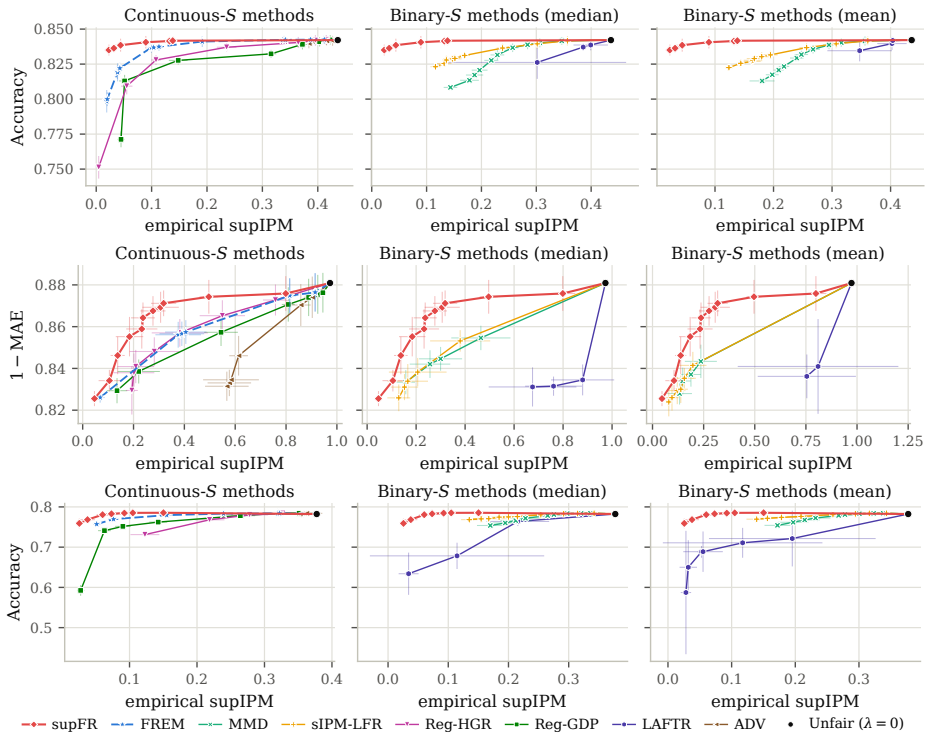


Figure 13: Prediction performance against the empirical supIPM of the representation on the test split.

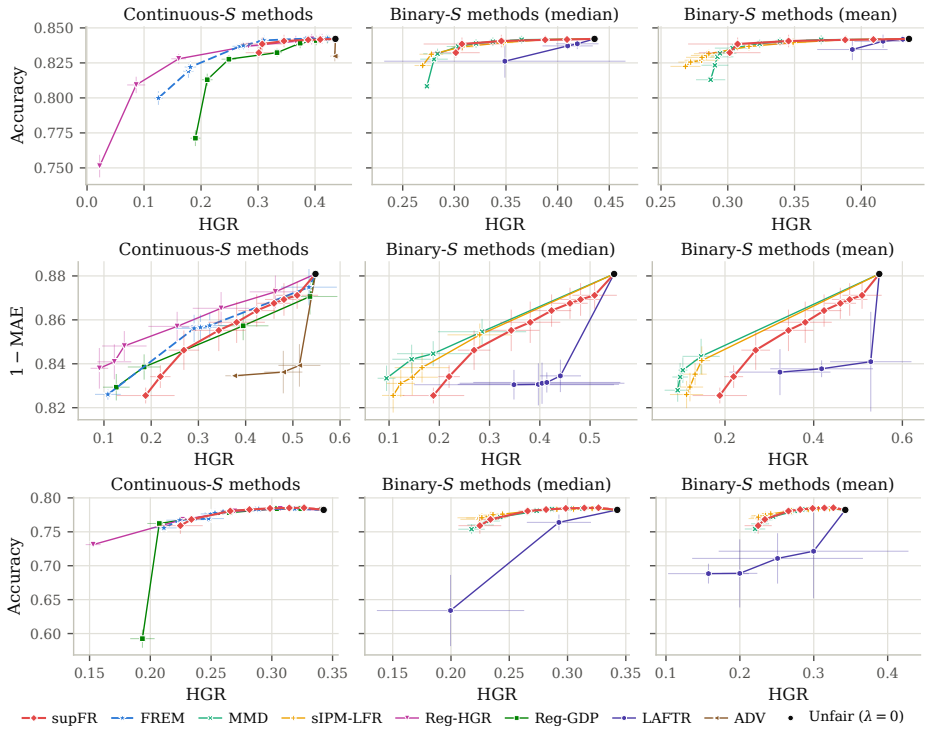


Figure 14: Prediction performance against the HGR maximal correlation between the fitted prediction and S.

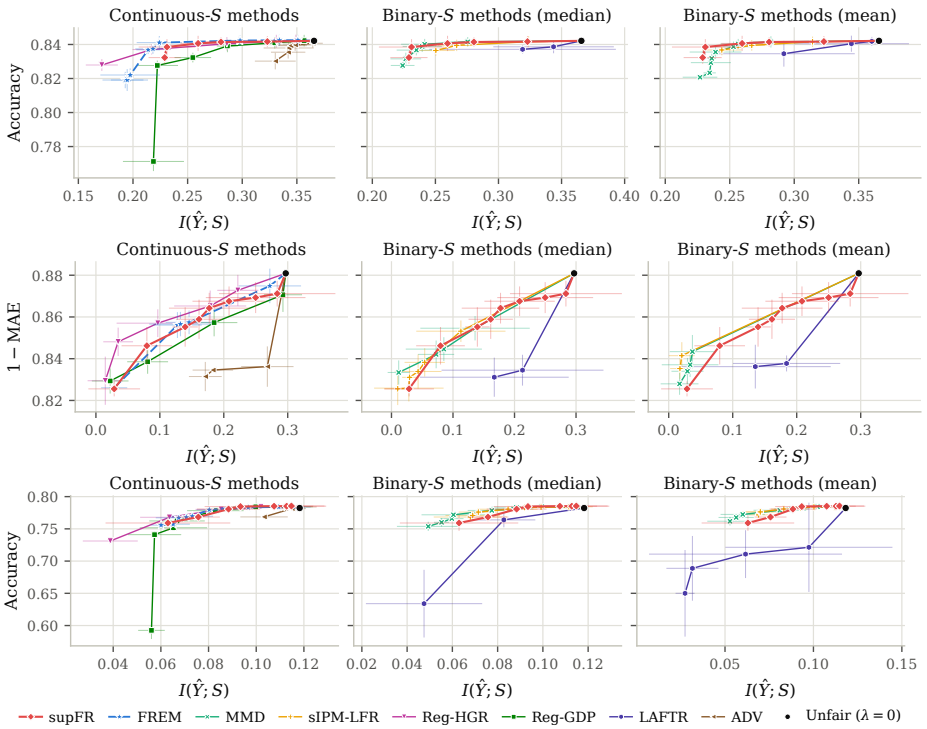


Figure 15: Prediction performance against the mutual information $I(\hat{Y}; S)$ between the fitted prediction and S.

Graph datasets. On Pokec-z and Pokec-n we compare supFR with FREM, Reg-GDP, Reg-HGR, ADV and MMD; LAFTR and sIPM-LFR are evaluated only on the tabular datasets. In Figures 16–18, the left column is Pokec-z and the right column is Pokec-n. In the top row of Figure 16, supFR attains a higher accuracy than FREM and Reg-GDP at a given prediction-level disparity on both graphs, and MMD and Reg-HGR reach a given disparity only at a lower accuracy. In the bottom row, supFR attains a smaller representation-level disparity than every baseline at comparable accuracy on both graphs.

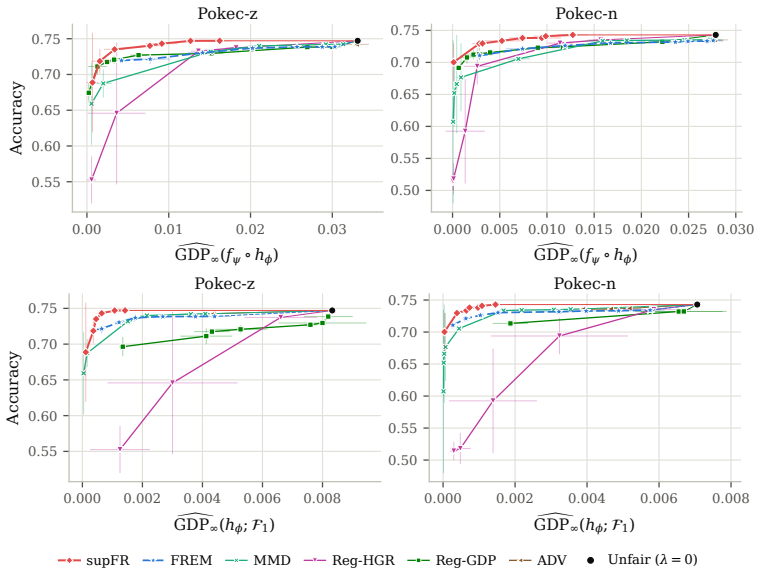


Figure 16: Prediction performance against the prediction-level disparity $\widehat{\mathrm{GDP}}_\infty(f_\psi \circ h_\phi)$ and the representation-level disparity $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$ on Pokec-z and Pokec-n.

Graph datasets under the other measures. Figures 17 and 18 compare the methods on the two graphs by the remaining measures of Appendix D.4, one measure per row. Under the kernel and the binned estimates of GDP, supFR attains the comparable results on both graphs. Under the empirical supIPM, its curve overlaps with that of FREM, and both lie well below Reg-GDP and Reg-HGR. Under the HGR maximal correlation and the mutual information, Reg-HGR attains the smallest values and supFR is comparable to FREM.

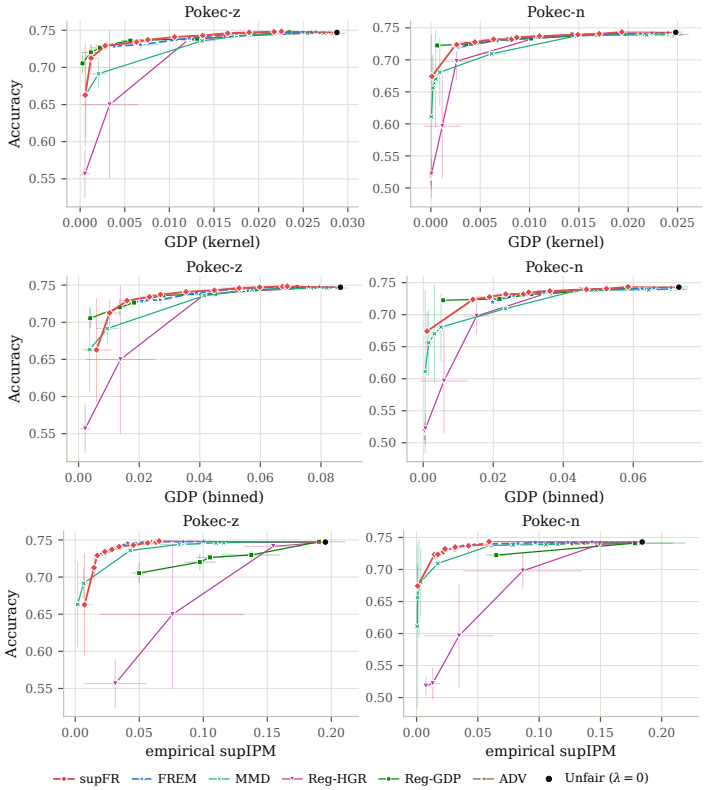


Figure 17: Prediction performance against the kernel GDP, the binned GDP and the empirical supIPM on Pokec-z and Pokec-n.

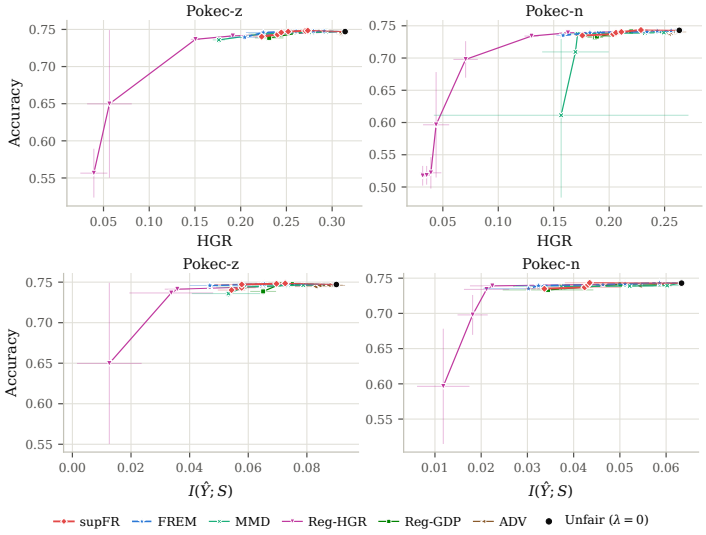


Figure 18: Prediction performance against the HGR maximal correlation and the mutual information $I(\hat{Y}; S)$ on Pokec-z and Pokec-n.

E COMPUTATIONAL COST

Appendix B.2 shows that, for fixed inner variables, the mini-batch objective of supFR and its gradient are unbiased for their counterparts on the whole training sample. The MMD penalty of FREM is biased on mini-batches, so it should, in principle, be computed on the whole training sample. Table 6 compares the cost of one gradient step on Adult for supFR with a mini-batch and for FREM and $\mathrm{supFR}_{\mathrm{MMD}}$ with a single batch of 32,561 observations, where $\mathrm{supFR}_{\mathrm{MMD}}$ denotes the variant of supFR in which the ReLU-IPM is replaced by the MMD.

Table 6: Computational cost on Adult.

Method	learning	seconds	peak memory (MB)
supFR	mini-batch (1,024)	0.086	33
FREM	batch (32,561)	out of memory	
$\mathrm{supFR}_{\mathrm{MMD}}$	batch (32,561)	out of memory	

The MMD evaluates the kernel at every pair of representations, so its cost grows quadratically with the number of observations, and on the whole training sample it exceeds the 24 GB of the GPU. FREM and $\mathrm{supFR}_{\mathrm{MMD}}$ are therefore trained with mini-batches, whose penalty is biased for the penalty on the whole sample, as noted in Appendix B.2. Figure 19 reports the time and the peak memory per step as the mini-batch grows, and the cross marks the first size that exceeds the memory of the GPU. FREM already needs 8.5 GB at 1,024 observations. The memory of supFR grows linearly and the whole training sample takes 129 MB, while its time per step is almost the same as with the mini-batch.

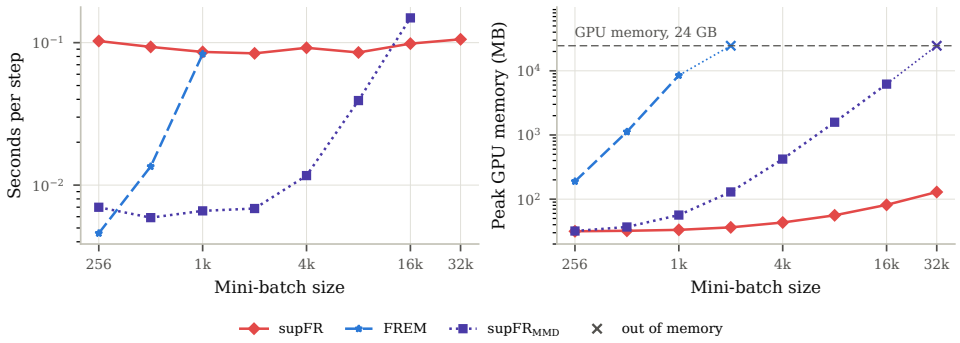


Figure 19: Computational cost on Adult as the mini-batch grows.

The linear cost of supFR, together with the unbiased mini-batch gradients of Appendix B.2, is the main reason for the ReLU discriminators of Section 3.2.

F EFFECT OF IPM CLASS

The discriminator class $\mathcal{V}$ determines which discrepancies between $\mathbb{P}_{\mathbf{Z}|S=s}$ and $\mathbb{P}_{\mathbf{Z}}$ the supremum in (9) detects, and hence which downstream heads the penalty controls. To examine its effect on supFR, we replace the ReLU class $\mathcal{V}_{\mathrm{R}}$ of (6) either by the sigmoid discriminators $\sigma(\theta^{\top}\mathbf{z}+\mu)$ of sIPM-LFR (Kim et al., 2022) or by the unit ball of the Gaussian RKHS, whose IPM is the MMD used by FREM (Kong et al., 2025); the latter variant is $\mathrm{supFR}_{\mathrm{MMD}}$ of Appendix E. Both variants retain the supremum over s in (9) and differ from supFR only in the discriminator class.

F.1 REAL DATA ANALYSIS

Figure 20 compares the three classes on the tabular datasets; the columns are Adult, Crime and ACSIncome from left to right, the top row reports the prediction-level disparity and the bottom row the representation-level disparity over $\mathcal{F}_1$. The architecture, the training settings of Appendix D.2

and the measures of Appendix D.4 are unchanged, and the penalty values are chosen separately for each discriminator class.

Figure 21 compares supFR, FREM and $\mathrm{supFR}_{\mathrm{MMD}}$ on Adult. Its left panel shows the test accuracy against the representation-level disparity over $\mathcal{F}_1$, and the circles mark the model of each method chosen at matched test accuracy as in Figure 5. Its right panel shows $\widehat{\Delta}_{\hat{f}_1}(s)$ of the three circled models over the sensitive domain.

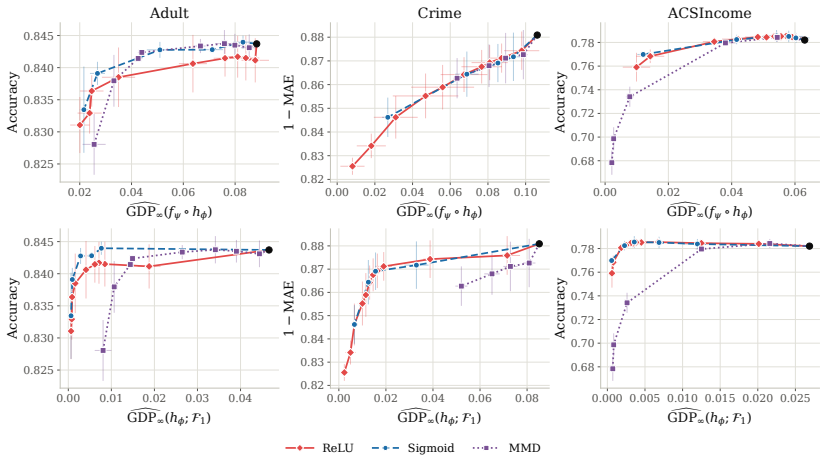


Figure 20: Prediction performance against the prediction-level disparity and the representation-level disparity over $\mathcal{F}_1$ for supFR trained with each discriminator class.

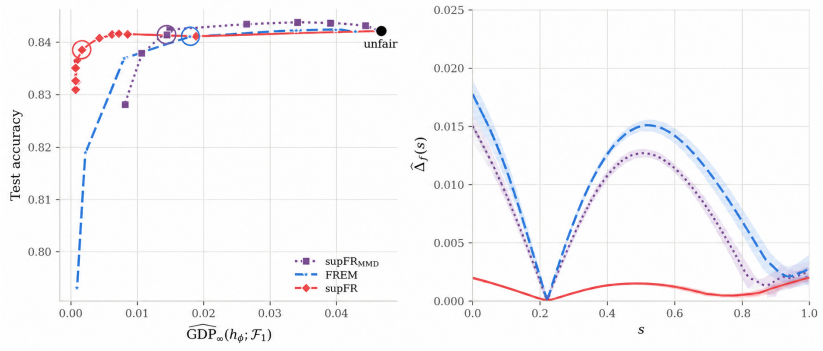


Figure 21: Left: test accuracy against $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$ on Adult for supFR, FREM and $\mathrm{supFR}_{\mathrm{MMD}}$. Right: $\widehat{\Delta}_{\hat{f}_1}(s)$ for the three models circled in the left panel, chosen at matched test accuracy.

Results. In Figure 20, the three discriminator classes yield similar prediction-level curves. The ReLU and the sigmoid discriminators attain comparable representation-level disparities, and their ordering varies with the dataset and with λ. Both attain a markedly lower representation-level disparity than the MMD at comparable prediction performance on the three datasets.

In the right panel of Figure 21, $\widehat{\Delta}_{\hat{f}_1}(s)$ of supFR is small over the whole sensitive domain, whereas that of $\mathrm{supFR}_{\mathrm{MMD}}$ is substantial near the left boundary and around $s \approx 0.5$, with a profile close to that of FREM. Since $\mathrm{supFR}_{\mathrm{MMD}}$ shares the discriminator class with FREM, this comparison suggests that the advantage of supFR over FREM in Figure 5 is attributable to the ReLU discriminator class. This observation is consistent with Section 3.2. The ReLU class controls the fairness violation of Hölder smooth prediction heads through (8), whereas the Gaussian RKHS unit ball contains only infinitely differentiable functions. These results do not undermine the motivation of supIPM. Computing the empirical EIPM with $\mathcal{V}_{\mathrm{R}}$ requires adversarial learning at each S_i , whereas the empirical supIPM in (13) requires a single maximization over (s, θ, μ), so the supremum over s makes the ReLU discriminator class computationally tractable for a continuous sensitive attribute.

F.2 Simulation Data Analysis

To further discuss Appendix F.1, we also compare supFR, $\mathrm{supFR}_{\mathrm{MMD}}$ and FREM on the simulation in Section 5.1, with the settings of Appendix C. Figure 22 shows the prediction performance against the representation-level disparity over $\mathcal{F}_1$ and $\widehat{\Delta}_{\hat{f}_1}(s)$ of the circled models, as in Figure 2.

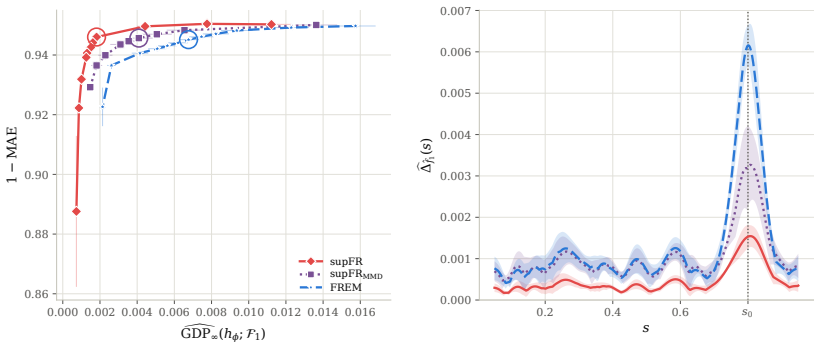


Figure 22: Left: prediction performance against $\widehat{\mathrm{GDP}}_\infty(h_\phi; \mathcal{F}_1)$ on simulated data for supFR, FREM and $\mathrm{supFR}_{\mathrm{MMD}}$. Right: $\widehat{\Delta}_{\hat{f}_1}(s)$ for the three models circled in the left panel.

Results. At comparable prediction performance, supFR attains the lowest representation-level disparity in the left panel, and its $\widehat{\Delta}_{\hat{f}_1}(s)$ is the smallest over the whole sensitive domain in the right panel. $\widehat{\Delta}_{\hat{f}_1}(s)$ of $\mathrm{supFR}_{\mathrm{MMD}}$ has $\widehat{\Delta}_{\hat{f}_1}$ close to that of FREM away from s_0, whereas at s_0 its peak is about half of that of FREM. Since the two share the discriminator class $\mathcal{V}$ and differ only in taking the supremum instead of the average over s, this difference is due to the supremum: for every $f \in \mathcal{V}$ and $s \in \mathcal{S}$,

$$\Delta_f(s) \leq \mathrm{IPM}_{\mathcal{V}}\left(\mathbb{P}_{\mathbf{Z}|S=s}, \mathbb{P}_{\mathbf{Z}}\right) \leq \mathrm{supIPM}_{\mathcal{V}}(\mathbf{Z}; S),$$

whereas the EIPM in (2) bounds only the average of $\Delta_f(S)$ over $\mathbb{P}_S$, to which the interval of small $\mathbb{P}_S$-probability around s_0 contributes little. Thus, the advantage of supFR over $\mathrm{supFR}_{\mathrm{MMD}}$ across the sensitive domain is attributable to the ReLU discriminator class, as in Appendix F.1, whereas the supremum over s further reduces the fairness violation at s_0.

G Ablation Studies

This appendix examines two choices in supFR on the tabular datasets: the kernel in the estimator of the distribution and the number T of candidate sensitive values. Each ablation varies one choice and keeps the architecture and the training settings of Appendix D.2 and the measures of Appendix D.4 unchanged. In each figure the columns are Adult, Crime and ACSIncome from left to right, the top row reports the prediction-level disparity and the bottom row the representation-level disparity over $\mathcal{F}_1$.

G.1 Kernel of Estimator

The kernel K_γ of (12) determines how the sensitive values neighboring s are weighted. Writing u for the distance between two sensitive values in units of the bandwidth, we compare the Gaussian kernel $\exp(-u^2/2)$ with the Epanechnikov kernel $(1-u^2)_+$, the triangular shape $(1-|u|)_+$ and the box kernel $\mathbf{1}\{|u| \leq 1\}$, the last three of which vanish beyond one bandwidth from s. The four are used at the same bandwidth.

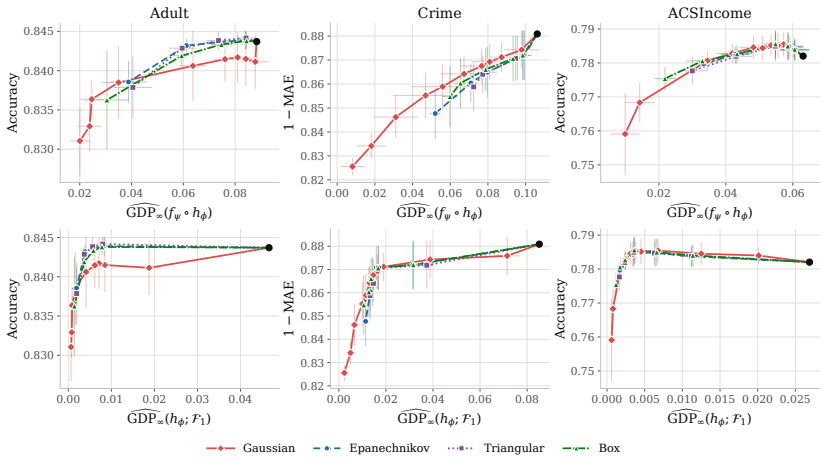


Figure 23: Prediction performance against the prediction-level disparity and the representation-level disparity over $\mathcal{F}_1$ for supFR trained with each smoothing kernel.

Results. In Figure 23, over the range of prediction performance shared by the four kernels, their curves are almost close on the three datasets, and no kernel attains the lowest disparity on every dataset. For the three compactly supported kernels, the prediction performance falls to that of a constant predictor at a smaller value of λ than for the Gaussian kernel, so their curves end at a larger disparity, whereas the curve of the Gaussian kernel extends to a lower disparity.

G.2 Number T of initial values of s.

The compass search over s of Appendix B.1 starts from T candidate sensitive values. We train supFR with $T \in \{5, 10, 15, 20, 25, 30\}$ in addition to $T = 17$ of Table 4, using the same random seeds and the same penalty values for every T.

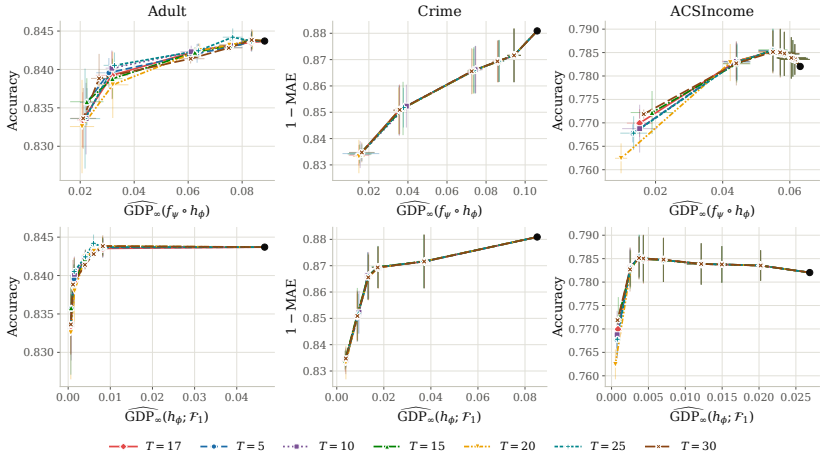


Figure 24: Prediction performance against the prediction-level disparity and the representation-level disparity over $\mathcal{F}_1$ for supFR trained at $T \in \{5, 10, 15, 17, 20, 25, 30\}$.

Results. In Figure 24, the seven values of T yield similar curves: on Crime the curves overlap, and on Adult and ACSIncome the differences between them lie within one standard deviation over the seeds and show no trend in T. Within the evaluated range of T, the results are therefore insensitive to the number of starting points of the compass search.