OPT2023: 15th Annual Workshop on Optimization for Machine Learning

f-FERM: A Scalable Framework for Robust Fair Empirical Risk Minimization

Sina Baharlouei BAHARLOU@USC.EDU
Shivam Patel SHIVAMAPATEL2002@GMAIL.COM
Meisam Razaviyayn RAZAVIYA@USC.EDU

Abstract

Numerous constraints and regularization terms have been proposed in the literature to promote fairness in machine learning tasks, most of these methods are not amenable to stochastic optimization due to the complex and nonlinear structure of constraints and regularizers. Here, the term "stochastic" refers to the ability of the algorithm to work with small mini-batches of data. Motivated by the limitation of existing literature, this paper presents a unified stochastic optimization framework for fair empirical risk minimization based on f-divergence measures (f-FERM). The proposed stochastic algorithm enjoys theoretical convergence guarantees. In addition, our experiments demonstrate the superiority of fairness-accuracy tradeoffs offered by f-FERM for almost all batch sizes (ranging from full-batch to batch size of one). Moreover, we show that our framework can be extended to the case where there is a distribution shift from training to the test data. Our extension is based on a distributionally robust optimization reformulation of f-FERM objective under ℓ_p norms as uncertainty sets. Again, in this distributionally robust setting, f-FERM enjoys not only theoretical convergence guarantees but also outperforms other baselines in the literature in the tasks involving distribution shifts. An efficient stochastic implementation of f-FERM is publicly available [1].

1. Introduction

Imposing statistical independence between model output and particular input features is of interest in various domains, especially when the generalization of a trained model is based on a collection of spurious features present in the training dataset [18, 23, 60]. These could be sensitive features like gender, race, age, and/or income in the context of fairness, or could be confounding factors like environmental artifacts in the context of image classification [4]. Existing literature on imposing statistical independence between selected input features and model outputs is directed into three approaches: pre-processing, post-processing, and in-processing methods. We reviewed the major advances in the three algorithmic fairness approaches in Appendix A.

This paper establishes a scalable (stochastic) fair empirical risk minimization framework through regularization via f-divergences (f-FERM) for both standard and distributed shift settings. f-FERM presents a unified methodology based on the Legendre-Fenchel transformation, enabling us to develop theoretically convergent first-order stochastic algorithms when only small batches of data are available at each iteration. Further, we have presented the first distributionally robust optimization framework under ℓ_p norms uncertainty sets covering nonconvex losses such as neural networks. The presented framework for fair inference in the presence of distribution shifts does not rely on the causal graph describing the causal interaction of input features, sensitive attributes, and target variables, which is rarely available in practical problems.

1. https://github.com/optimization-for-data-driven-science/f-FERM

2. Fair Empirical Risk Minimization via f-divergences

A widely studied problem in algorithmic fairness is promoting a notion of group fairness, such as demographic parity, equalized odds, equality of opportunity, or sufficiency through an in-processing method. For these notions, we aim to establish a [conditional] statistical independence between the predictions (e.g., the creditworthiness of the individual) and the sensitive attributes (e.g., gender, race). For simplicity of presentation, we formulate all problems under the demographic parity notion, which requires statistical independence between the prediction and the sensitive attribute. Without loss of generality, all formulations and methods are generalizable to other aforementioned notions of group fairness by considering conditional random variables (see Appendix B). A popular in-processing approach for training fair (classification) models under the demographic parity notion is to regularize the empirical risk minimization:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \mathcal{D}\Big(\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s), \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x})) \otimes \mathbb{P}(s)\Big), \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter to be learned (e.g. weights of the neural network); $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th input feature vector; y_i is the actual label/class for sample i; $\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i)$ is the prediction of the model for sample i; and $\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$ is the loss function measuring the "goodness-of-fit" for sample i. Here, $\mathcal{D}$ is a divergence between the joint probability distribution of the predictions and sensitive attributes and the Kronecker product of their marginal distributions. Recall that $\hat{y}_{\boldsymbol{\theta}}$ and s are statistically independent iff $\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s)$ follows $\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x})) \otimes \mathbb{P}(s)$. Therefore, the second term in (1) is zero if and only if $\hat{y}_{\boldsymbol{\theta}}$ and s are statistically independent (complete fairness under the demographic parity notion).

This section studies the fair empirical risk minimization regularized by a broad class of f-divergence measures. Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete probability measures taking values in $\mathcal{P} = \{1, \ldots, m\}$. The f-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is defined as [43, Def 4.9](see Appendix C for the general continuous case):

$$\mathcal{D}_f(\mathbb{P}, \mathbb{Q}) = \sum_{j=1}^{m} \mathbb{Q}_j f\Big(\frac{\mathbb{P}_j}{\mathbb{Q}_j}\Big) \tag{2}$$

The above definition covers many well-known divergence measures (used for imposing fairness), such as KL-divergence for the choice of $f(t) = t\log(t)$ [50], or χ^2 divergence when $f(t) = (t-1)^2$ [35]. As shown in Appendix D, $\mathcal{D}_f$ in (1) is zero *if and only if* the probability distribution of s and $\hat{y}_{\boldsymbol{\theta}}$ *are statistically independent* for the choices of f listed in Table 1. In addition, we prove that these f-divergences either cover or provide upper bounds for the popular notions of fairness violations in the literature, such as ℓ_p distances, Rényi correlation [5], and demographic parity (equalized odds) violation. Further, unlike Rényi correlation [5, 22], we can utilize Legendre-Fenchel duality (and variational representation) to develop (provably) convergent algorithms with **stochastic (mini-batch) updates**. The stochastic optimization method for is described in the next subsection.

2.1. A Convergent Stochastic Algorithm for fair ERM via f-Divergences

Let us start by rewriting (1) using f-divergences as the divergence measure:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}} \mathbb{P}_s(s=k)\mathbb{P}_{\hat{y}_{\boldsymbol{\theta}}}(\hat{y}_{\boldsymbol{\theta}}) f\Big(\frac{\mathbb{P}_{\hat{y}_{\boldsymbol{\theta}},s}(\hat{y}_{\boldsymbol{\theta}}=j, s=k)}{\mathbb{P}_{\hat{y}_{\boldsymbol{\theta}}}(\hat{y}_{\boldsymbol{\theta}}=j)\mathbb{P}_s(s=k)}\Big) \tag{f-FERM}$$

In particular, directly evaluating the gradient of the objective function of (f-FERM) on a mini-batch of data leads to a statistically biased estimation of the entire objective's gradient. Such statistical

biases prevent the convergence of algorithms such as SGD (even with a strongly convex minimization landscape) [2, 10], let aside the more complex objectives arising in modern-day neural networks.

To derive stochastic algorithms, one can use the variational forms of f-divergences to delineate them as a pointwise supremum of affine transformation over probability densities. The most commonly used and well-behaved transform is the Legendre-Fenchel transform (often called the convex conjugates), which linearizes the dependence of the objective function to input data points using a variational reformulation. Particularly, we can rewrite (f-FERM) using the following result:

Proposition 1 *Let $f(\cdot)$ be a convex function. Then, (f-FERM) can be reformulated as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \quad \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}} \Big[\mathbf{A}_{jk}\mathbb{P}_{\hat{y},s}(\hat{y}_{\boldsymbol{\theta}} = j, s = k) - f^*(\mathbf{A}_{jk})\mathbb{P}_{\hat{y}}(\hat{y}_{\boldsymbol{\theta}} = j)\mathbb{P}_s(s = k)\Big] \tag{3}$$

where $f^*(z) = \sup_{w\in\text{dom}(f)} w^T z - f(w)$ is the Legendre-Fenchel transformation of the function f.

Proof The proof is standard and appears in Appendix E. ■

As a result, Problem (3) can be written as a linearly separable function of input data points ($\mathbf{x}_i$'s):

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \quad \frac{1}{n}\sum_{i=1}^{n}\left[\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}} \Big[\mathbf{A}_{jk}F_j(\mathbf{x}_i;\boldsymbol{\theta})\mathbb{1}(s_i = k) - f^*(\mathbf{A}_{jk})\pi_k F_j(\mathbf{x}_i;\boldsymbol{\theta})\Big]\right] \tag{4}$$

Where $\pi_k := \mathbb{P}_s(s = k) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(s_i = k)$ and $F_j(\mathbf{x}_i;\boldsymbol{\theta})$ is the j-th entry of the softmax layer output for datapoint $\mathbf{x}_i$. Therefore, evaluating the gradient of the objective function with respect to the optimization variables $\boldsymbol{\theta}$ and $\mathbf{A}$ over a random small batch of data points leads to an unbiased gradient estimator of the entire objective function. In addition to providing an unbiased estimator of gradients, the reformulation (4) has another crucial property: *the objective function is concave in* $\mathbf{A}$. Therefore, optimization problem (4) falls under the category of nonconvex-concave min-max optimization problems. That is, the objective is (possibly) nonconvex in $\boldsymbol{\theta}$ and is concave in $\mathbf{A}$. Thus, we can borrow tools from the (stochastic) nonconvex-concave min-max optimization literature [32, 34, 46] to derive a convergent first-order stochastic algorithm as presented in Algorithm 1. We list the closed-form of $f(\cdot)$, $f^*(\cdot)$, for several widely-used f-divergence measures in Table 1. For the derivation, see Appendix F.

Algorithm 1 Stochastic Gradient Descent-Ascent (SGDA) for f-FERM

1: **Input**: $\boldsymbol{\theta}^0 \in \mathbb{R}^{d_\theta}$, step-sizes $\eta_{\boldsymbol{\theta}}$, η_α, fairness parameter $\lambda \geq 0$, iteration number T, Batchsize b
2: **for** $t = 1, \ldots, T$ **do**
3: Sample minibatch of data $\mathcal{B}_t = \{(\mathbf{x}_{t1}, \mathbf{y}_{t1}), \cdots, (\mathbf{x}_{tb}, \mathbf{y}_{tb})\}$
4: $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \frac{\eta_{\boldsymbol{\theta}}}{b}\sum \nabla_{\boldsymbol{\theta}}\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), y) - \eta_{\boldsymbol{\theta}}\lambda\nabla_{\boldsymbol{\theta}}\Big(\mathbf{A}_{jk}^{t-1}\hat{\mathbb{P}}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}},\mathbf{s}}(j,k;\,\mathcal{B}_t) - \pi_k f^*(\mathbf{A}_{jk}^{t-1})\hat{\mathbb{P}}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}(j;\,\mathcal{B}_t)\Big)$
5: $\mathbf{A}_{jk}^{t} = \mathbf{A}_{jk}^{t-1} + \eta_\alpha\,\nabla_{\mathbf{A}}\Big(\mathbf{A}_{jk}^{t-1}\hat{\mathbb{P}}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}},\mathbf{s}}(j,k;\,\mathcal{B}_t) - \pi_k f^*(\mathbf{A}_{jk}^{t-1})\hat{\mathbb{P}}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}(j;\,\mathcal{B}_t)\Big)$
6: **Return:** $\boldsymbol{\theta}^T$

Theorem 2 ***(Informal Statement)*** *Assume that $\ell(\cdot,\cdot)$ and $\mathcal{F}_j(\cdot,\boldsymbol{\theta})$ are Lipschitz continuous for any given j and $\boldsymbol{\theta}$ and their gradients are L-Lipshitz. Further, assume that $\mathbb{P}(s = k) > 0$ for all protected groups and $\mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) > 0$ at every iteration for all labels j. Then, for any given batch size $1 \leq |\mathcal{B}| \leq n$, Algorithm 1 finds an ϵ-stationary solution of (f-FERM) in $\mathcal{O}(\frac{1}{\epsilon^8})$ for any given $\epsilon > 0$.*

Table 1: Unbiased Estimators for f-divergence Regularizers

Divergence	$f(t)$	The term r_{jk} inside regularizer $\lambda \sum_{j,k} r_{jk}$ in (4)
χ^2	$(t-1)^2$	$\pi_k[\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} - (\mathbf{A}_{jk} + \frac{\mathbf{A}_{jk}^2}{4})\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]$
Reverse KL	$-\ln t$	$\pi_k[\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} + (1+\ln(-\mathbf{A}_{jk}))\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]$
Total Variational	$\frac{1}{2}\|t-1\|$	$\pi_k\mathbf{A}_{jk}[\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} - \mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]\mathbb{I}_{\{\|\mathbf{A}_{jk}\|<1/2\}}$
KL	$t\ln t$	$\pi_k[\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} - e^{\mathbf{A}_{jk}-1}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]$
Jensen-Shannon	$-(t+1)\ln(\frac{t+1}{2}) + t\ln t$	$\pi_k[\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} + \ln(2-e^{\mathbf{A}_{jk}})\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]$
Squared Hellinger	$(\sqrt{t}-1)^2$	$\pi_k[\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}\|\mathbf{s}_k} + (\mathbf{A}_{jk}^{-1}+2)\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}]$

Proof The formal statement and proof are relegated to Appendix G. ■

Theorem 2 applies to all f-divergences listed in Table 1 for all batch-sizes (even as small as the batch size of 1). More sophisticated algorithms can be used to obtain $\mathcal{O}(\epsilon^{-6})$ iteration complexity [45, 63]. However, such algorithms use nested loops and require more hyperparameter tunings. If the f-divergence leads to a strongly concave function in $\mathbf{A}$ or satisfies Polyak-Łojasiewicz condition (e.g., for χ^2 divergence), a faster rate of $\mathcal{O}(\epsilon^{-5})$ can be obtained for this algorithm (Appendix G). In addition, if larger batch size of $\mathcal{O}(\epsilon^{-2})$ is used, we can further improve this rate to $O(\epsilon^{-4})$ iteration complexity (see Appendix G). Finally, when full batch size is used, then double/triple-loop algorithms can lead to the iteration complexity bounds of $O(\epsilon^{-2})$ in the nonconvex-strongly concave setting and $O(\epsilon^{-3})$ in the general nonconvex-concave setting; see [28, 40, 42, 55].

3. Robust f-FERM in the Presence of Distribution Shifts

In the previous section, we assumed that the training and test domains have the same distribution. However, this assumption is not necessarily valid in certain applications [19]. In particular, a model that behaves fairly on the training data distribution may have an unfair performance in the test phase. To address this issue, this section develops stochastic algorithms for fair empirical risk minimization via f-divergences in the presence of the distribution shifts. Assume that $\hat{\mathbb{P}}_{s,y}(s,\hat{y})$ is the joint distribution of sensitive attributes and predictions on the training data. The distributionally robust fair empirical risk minimization via f-divergences is formulated as:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) \quad \text{s.t.} \quad \max_{\mathbb{P}\in\mathcal{B}} \mathcal{D}_f\Big(\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s) \| \, \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x})) \otimes \mathbb{P}(s)\Big) \leq \delta. \tag{5}$$

Here $\mathcal{B}$ is the distributional uncertainty set. This section focuses on the widely studied ℓ_p norms as the uncertainty set for the distributional distance between the training and test domains. In this case, Problem (5) can be written as:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) \quad \text{s.t.} \quad \max_{\substack{\|\mathbb{P}-\hat{\mathbb{P}}\|_p\leq\delta \\ \|\mathbb{Q}-\hat{\mathbb{Q}}\|_p\leq\delta}} \mathcal{D}_f(\mathbb{P}\|\mathbb{Q}) \leq \kappa, \tag{6}$$

where $\hat{\mathbb{P}}$ represents the joint distribution of the sensitive attributes and predictions and $\hat{\mathbb{Q}}$ denotes the Kronecker product of the marginal distributions between sensitive attributes and predictions. Since handling non-convex constraints is challenging, as it is standard in training machine learning models, we consider the Lagrangian relaxation of Problem (6) as follows:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \max_{\substack{\|\mathbb{P}-\hat{\mathbb{P}}\|_p\leq\delta \\ \|\mathbb{Q}-\hat{\mathbb{Q}}\|_p\leq\delta}} \mathcal{D}_f(\mathbb{P}\|\mathbb{Q}) \tag{7}$$

This problem falls under the nonconvex-nonconcave, min-max optimization category and is computationally intractable for general uncertainty sets [13]. However, such a min-max optimization problem can be solved to stationarity when the diameter of set $\mathcal{B}$ is small (i.e., under small domain shift), see [41]. The core idea is to approximate the inner maximization problem with the Taylor approximation, leading to a nonconvex-concave min-max optimization, which is easier to solve [13, 47]. This idea has been used and been successful in machine learning (see Foret et al. [20] for its use in Sharpness-aware minimization). Utilizing this idea, Problem (7) can be approximated as:

$$\min_{\boldsymbol{\theta}} \max_{\substack{\|\mathbb{U}\|_p \leq \delta \\ \|\mathbb{V}\|_p \leq \delta}} \left(h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V}) := \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \langle \mathbb{U}, \nabla_{\mathbb{P}} \mathcal{D}_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}}) \rangle + \lambda \langle \mathbb{V}, \nabla_{\mathbb{Q}} \mathcal{D}_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}}) \rangle \right), \tag{8}$$

where we used the change of variables $\mathbb{U} := \mathbb{P} - \hat{\mathbb{P}}$ and $\mathbb{V} := \mathbb{Q} - \hat{\mathbb{Q}}$. Equivalently,

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda\delta \|\nabla_{\mathbb{P}} \mathcal{D}_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}})\|_q + \lambda\delta \|\nabla_{\mathbb{Q}} \mathcal{D}_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}})\|_q, \tag{9}$$

where $\|\cdot\|_q$ is the dual of the ℓ_p norm with $\frac{1}{p} + \frac{1}{q} = 1$. It can be shown that finding stationary points of (9) leads to a stationary point of (6):

Proposition 3 *Assume that the gradient of the loss function is L-Lipshitz, and the second-order derivative of the loss exists. Then, a given $\epsilon-$approximate stationary solution of Problem (9) is an $O(\epsilon)-$approximate stationary solution of Problem (7) whenever $L\delta \lesssim \epsilon$.*

This proposition is an immediate application of Ostrovskii et al. [41, Theorem 3.1], To solve the problem, we need to obtain the (sub)-gradients of the objective function in (8) w.r.t the $\boldsymbol{\theta}$, $\mathbb{U}$, and $\mathbb{V}$ variables. First, notice that

$$\nabla_{\mathbb{U}} h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V}) = \nabla_{\mathbb{P}} D_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}}) = \boldsymbol{\alpha}^*(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) \text{ and } \nabla_{\mathbb{V}} h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V}) = \nabla_{\mathbb{Q}} D_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}}) = f^*(\boldsymbol{\alpha}^*(\hat{\mathbb{P}}, \hat{\mathbb{Q}})),$$

where $\boldsymbol{\alpha}^*(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) \in \arg\max_{\boldsymbol{\alpha}} \sum_j \alpha_j \hat{p}_j(\boldsymbol{\theta}) - \hat{q}_j(\boldsymbol{\theta}) f^*(\alpha_j)$. Here we invoked Danskin's theorem on the variational form of D_f; $\hat{p}_j(\boldsymbol{\theta})$ and $\hat{q}_j(\boldsymbol{\theta})$ is the j-th element of $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$, respectively. Next, we need to compute $\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V})$. Notice that the derivative of the first term in $h(\cdot)$ w.r.t. $\boldsymbol{\theta}$ is easy to compute. We next calculate the derivative of the second term of $h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V})$ w.r.t. $\boldsymbol{\theta}$. As the derivative of the third term can be computed similarly, we omit its derivation here.

$$\nabla_{\boldsymbol{\theta}} \langle \mathbb{U}, \nabla_{\mathbb{P}} \mathcal{D}_f(\hat{\mathbb{P}} || \hat{\mathbb{Q}}) \rangle = \nabla_{\boldsymbol{\theta}} \langle \mathbb{U}, \boldsymbol{\alpha}^*(\hat{\mathbb{P}}, \hat{\mathbb{Q}}) \rangle = \sum_j u_j \frac{\hat{q}_j(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \hat{p}_j(\boldsymbol{\theta}) - \hat{p}_j(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \hat{q}_j(\boldsymbol{\theta})}{\hat{q}_j^2(\boldsymbol{\theta}) \times (f^*)''(\alpha)|_{\alpha = \alpha_j^*(\hat{\mathbb{P}}, \hat{\mathbb{Q}})}} \tag{10}$$

where in the last equation, we used the implicit function theorem to compute the derivative of $\boldsymbol{\alpha}^*$ w.r.t. $\boldsymbol{\theta}$. Notice that an implicit assumption here is that f is differentiable (which holds for KL-divergence, χ^2 divergence, reverse KL, Jensen-Shannon, and Squared Hellinger distance). Having access to the gradients, we can apply the standard [sub-]gradient descent-ascent algorithm to obtain a solution to Problem (9). We proposed a semi-stochastic algorithm for solving Problem (9) in Appendix H. Further, we study the fair inference in the presence of the distribution shift when ϵ is large.

4. Experiments

This section evaluates the performance and efficiency of the proposed f-divergence frameworks in Section 2 through extensive experiments on benchmark datasets against several state-of-the-art

approaches. We use 3 popular notions of group fairness, including demographic parity, equalized odds, and equality of opportunity violations (see Appendix B for the exact definitions) to measure the fairness of trained models. To run Algorithm 1, we set $\eta_{\boldsymbol{\theta}}$ and η_α to 10^{-5} and 10^{-6} respectively in all experiments. Further, by changing λ, we get different points in the trade-off curve between the accuracy and fairness of the trained model. The range of values for λ depends on the f-divergence we use (see Appendix I for more information on tuning hyper-parameters).

In the first set of experiments, we compare different f-divergence formulations for (f-FERM) to each other and several state-of-the-art approaches supporting multiple sensitive attributes. Figure 1 demonstrates the given tradeoff on the adult dataset [7] with gender and race as the sensitive attributes (black-female, black-male, white-female, white-male). To measure fairness, we use the demographic parity violation defined as:

$$\text{DPV} = \max_{i,j \in \mathcal{S}} |\mathbb{P}(\hat{y} = 1|s = i) - \mathbb{P}(\hat{y} = 1|s = j)|$$

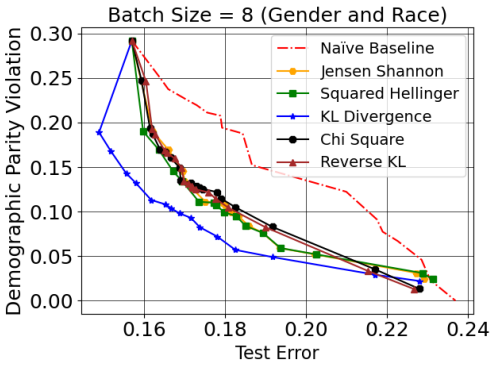

Figure 1: Performance of different f-divergences as the regularizers for measuring fainress violation. The experiment is on the adult dataset with gender and race as sensitive attributes. While the offered tradeoffs are close to each other for small demographic parity violations, KL-divergence shows an extraordinary performance for a low-fairness high-accuracy regime. We do not display the performance for larger batch sizes or when only one sensitive attribute is available due to the insignificant difference between the performance of different f-divergences.

In the case of binary sensitive attributes (e.g., gender), there is no significant variation between different f-divergences. However, when we have 2 sensitive attributes and the batch size is small (8 in Figure 1), the results significantly differ for various f-divergences. Further, in Figure 2, we compare one of the f-divergences (reverse KL) to several SOTA methods including Baharlouei et al. [5], Cho et al. [11], Mary et al. [38]. Other approaches such as the pre-processing method of Zemel et al. [62], post-processing approach of Hardt et al. [23], and several in-processing methods including Donini et al. [16], Jiang et al. [26], Zafar et al. [61] demonstrate lower performance compared to the ones depicted in Figure 2 and are removed from the figure. While our approach demonstrates consistently good performance across different batch sizes (full-batch, 64, 8, 2), other methods' performances drop significantly for smaller ones. For further experiments on other datasets (German and COMPAS) and other fairness measures (equality of opportunity and equalized odds violations), see Appendix J.

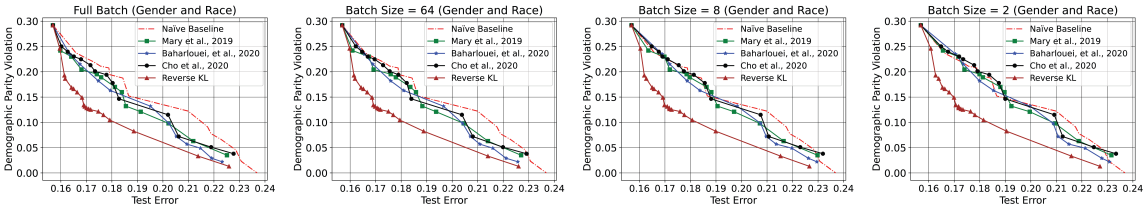

Figure 2: Performance of the trained fair models on Adult Dataset with gender and race as two sensitive attributes with different Batch-sizes. The red dashed line represents the Naïve baseline where the model outputs zero with probability p. By increasing p, the model becomes fairer at the cost of the loss in accuracy.

References

[1] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1418–1426, 2019.

[2] Ahmad Ajalloeian and Sebastian U. Stich. Analysis of SGD with biased gradient estimators. *CoRR*, abs/2008.00051, 2020. URL https://arxiv.org/abs/2008.00051.

[3] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

[6] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

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

[8] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[9] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

[10] Jie Chen and Ronny Luss. Stochastic gradient descent with biased but consistent gradient estimators. *CoRR*, abs/1807.11880, 2018. URL http://arxiv.org/abs/1807.11880.

[11] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099, 2020.

[12] Jessica Dai and Sarah M Brown. Label bias, label shift: Fair machine learning with unreliable labels. In *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*, volume 12, 2020.

[13] Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

[14] Yuyang Deng, Mohammad Mahdi Kamani, Pouria Mahdavinia, and Mehrdad Mahdavi. Distributed personalized empirical risk minimization. In *International Workshop on Federated Learning for Distributed Data Mining*, 2023.

[15] 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.

[16] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[17] Wei Du and Xintao Wu. Fair and robust classification under sample selection bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2999–3003, 2021.

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

[19] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 33:11996–12007, 2020.

[20] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[21] Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip S Thomas, and Scott Niekum. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2021.

[22] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural rényi minimization for continuous features. In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, pages 2262–2268. International Joint Conferences on Artificial Intelligence Organization, 2020.

[23] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[24] Hisham Husain. Distributional robustness with ipms and links to regularization and gans. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11816–11827. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8929c70f8d710e412d38da624b21c3c8-Paper.pdf.

[25] Aleksandr Davidovich Ioffe and Vladimir Mihajlovič Tihomirov. *Theory of extremal problems*. Elsevier, 2009.

[26] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR, 2020.

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

[28] Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

[29] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[30] Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. Impossibility results for fair representations. *arXiv preprint arXiv:2107.03483*, 2021.

[31] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

[32] Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth nonconvex-nonconcave minimax optimization: Primal-dual balancing and iteration complexity analysis, 2023.

[33] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

[34] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6083–6093. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lin20a.html.

[35] Andrew Lowy, Sina Baharlouei, Rakesh Pavan, Meisam Razaviyayn, and Ahmad Beirami. A stochastic optimization framework for fair risk minimization. *tmlr*, 2022.

[36] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? *Advances in Neural Information Processing Systems*, 34:25773–25784, 2021.

[37] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023.

[38] Jeremie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mary19a.html.

[39] Alan Mishler and Niccolò Dalmasso. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. *arXiv preprint arXiv:2202.05049*, 2022.

[40] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[41] Dmitrii M Ostrovskii, Babak Barazandeh, and Meisam Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, 2021.

[42] Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

[43] Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning. *Book draft*, 2022.

[44] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

[45] H Rafique, M Liu, Q Lin, and T Yang. Non-convex min–max optimization: provable algorithms and applications in machine learning (2018). *arXiv preprint arXiv:1810.02060*, 1810.

[46] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020. doi: 10.1109/MSP.2020.3003851.

[47] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[48] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9419–9427, 2021.

[49] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

[50] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022.

[51] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[52] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[53] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1770ae9e1b6bc9f5fd2841f141557ffb-Paper.pdf.

[54] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification, 2020.

[55] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.

[57] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey. *CoRR*, abs/2111.03015, 2021. URL https://arxiv.org/abs/2111.03015.

[58] Haotao Wang, Junyuan Hong, Jiayu Zhou, and Zhangyang Wang. How robust is your fairness? evaluating and sustaining fairness under unseen distribution shifts. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=11pGlecTz2.

[59] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33:5190–5203, 2020.

[60] Ke Yan, Lu Kou, and David Zhang. Learning domain-invariant subspace using domain features and independence maximization. *IEEE transactions on cybernetics*, 48(1):288–299, 2017.

[61] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

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

[63] Xuan Zhang, Necdet Serhat Aybat, and Mert Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *Advances in Neural Information Processing Systems*, 35:21668–21681, 2022.

[64] Meiyu Zhong and Ravi Tandon. Learning fair classifiers via min-max f-divergence regularization, 2023.

Appendix A. A Review of Algorithmic Fairness Methods

Pre-processing methods entail upstream changes made in datasets to mask sensitive features or reduce the dependency of output variables on sensitive features through transforming data in a stage before the training phase [27, 56, 62]. Post-processing methods involve model-specific adjustments to the model's output to ensure the independence of predictions and sensitive attributes [3, 23]. While pre-processing and post-processing methods do not affect the training procedure, they fail to exploit underlying training mechanisms for the best achievable accuracy-fairness tradeoffs. Unsurprisingly enough, optimizing accuracy and fairness jointly (in-processing) leads to better tradeoffs than sequentially optimizing fairness and accuracy in a pre-processing or post-processing fashion.

In-processing methods overcome these shortcomings of pre-processing and post-processing approaches by adding fairness constraints or regularizers to the training objective, penalizing dependence between sensitive attributes and output variables. [61] utilizes covariance as the measure of independence between the sensitive attributes and the predictions. While such a measure is amenable to stochastic updates, it fails to capture correlations beyond linear. Alternatively, several non-linear measures such as Rényi correlation [5], χ^2 divergence [35], L_∞ distance [16], and Maximum Mean Discrepancy (MMD) [44] are proposed in the literature to establish the independence of the predictors and sensitive attributes. In-processing techniques can be model-specific [1, 57] or generalizable to different training algorithms [5, 35].

In the spirit of in-processing methods, input data-driven constraints or regularization terms are used to modify training objectives of problems like learning generalizable models to new environments, invariant learning, and learning in the presence of distribution shifts [4, 5, 38]. Such constrained/regularized reformulations are prevalent in learning robust classifiers against adversarial attacks [52], meta-learning [6], federated learning [14], and alternative learning paradigms such as learning distributionally robust optimization (DRO) models [29, 31], tilted empirical risk minimization (TERM) [33], and Squared-root Lasso [8].

While in-processing techniques outperform pre-processing and post-processing approaches, they are not scalable to large datasets because of a lack of adaptability to Stochastic Approximation methods [35, 38]. From an optimization perspective, all aforementioned examples consist of regularization terms in their objective functions where the gradient cannot be described as a linear combination of data points functions. As a result, applying stochastic gradient descent or other stochastic first-order methods on the objective functions of such problems might not converge, especially for small batch sizes.

Motivated by this, [35] proposes a provably convergent stochastic optimization framework for Exponential Rényi Mutual Information as the measure of independence. More recently Zhong and Tandon [64] use f-divergences as regularization terms to establish the independence between sensitive attributes and predictions. They estimate the f-divergence regularizers offline through multi-layer neural networks to avoid the computational challenges of devising scalable stochastic methods for nonconvex min-max problems. Our approach, on the other hand, directly solves the variational formulation for both full-batch and stochastic settings with convergence guarantees to non-spurious solutions. In Section 2, using the variational representation of f-divergences, we present a convergent stochastic optimization framework for fair learning via f-divergences. [35] is a special case of f-divergences where $f(t) = t^2 - 1$ (χ^2 divergence). Aside from χ^2, all other divergences listed in Table 1 are not introduced in the literature to the best of our knowledge.

The need to have convergent algorithms for fair empirical risk minimization does not end with designing methods amenable to stochastic approximation. Detection and mitigation of biases against protected groups in the presence of distribution shifts have been extensively studied in recent years. Lechner et al. [30] theoretically shows that learning fair representations (pre-processing) is nearly *impossible* for the popular notions of fairness, such as demographic parity in the presence of the distribution shift. Ding et al. [15], on the other hand, experimentally demonstrates that applying post-processing fairness techniques [23] to learn fair predictors of income concerning race, gender, and age fails to transfer from one US state (training domain) to another state. Overlooking distribution shifts can lead to catastrophic decisions threatening the well-being of human subjects when deploying a trained model in certain hospitals to other hospitals [49]. The current literature for handling distribution shifts with in-processing methods relies on certain assumptions on the type of distribution shift (demographic shift [17, 19, 21, 36], label shift [12], and/or covariate shift [48, 51]) or explicit access to the **causal graph** [39, 49] of predictors, sensitive attributes, and target variables. As a result, they face practical limitations and cannot cope with most real-world problems involving complex shifts that cannot be categorized in the ones assumed in their works.

Alternatively, [54] provides convex objective functions for imposing fairness on logistic regression using constraint optimization. [53] use MMD for defining uncertainty sets around training distribution, whereas [24] use Integral Probability Measure (IPM) to mitigate the distribution shift. The main limitation of these approaches is their reliance on the convexity of the underlying learning model and lack of scalability due to incompatibility with stochastic optimization algorithms. Wang et al. [58] uses the Maximum Mean Discrepancy (MMD) distance between the spectral norm of the Hessian matrix at advantaged and disadvantaged data points. However, they do not provide convergence guarantees for their proposed algorithm to any notion of optimality. In addition, the method is not necessarily amenable to stochastic updates. While we naturally define the uncertainty set directly on the joint distribution of sensitive attributes and predictions, they use the curvature of the obtained solution quantified by the norm of the Hessian matrix as a heuristic for promoting the robustness of the fair solution. In Appendix H, we present our approach for fair inference in the presence of the distribution shifts in detail.

Appendix B. f-FERM for Other Notions of Group Fairness

This section shows how we can use alternative notions of fairness, such as equality of opportunity of equalized odds [23] instead of demographic parity violation in f-FERM.

Note that a trained model satisfies the equality of opportunity notion for a given binary classifier with a binary sensitive attribute if and only if:

$$\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}) = 1, s = i |\, y = 1) = \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}) = 1, s = j |\, y = 1) \quad \forall\, i, j \in \mathcal{S} \tag{11}$$

Therefore, to have a framework for fair inference via f-divergences under the equality of opportunity notion, we optimize:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \mathcal{D}_f\Big(\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s | y = 1) || \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}) | y = 1) \otimes \mathbb{P}(s | y = 1)\Big). \tag{12}$$

Practically, it means that for evaluating the probability measures in the regularization term, we need to focus on the data points whose target labels are 1.

Further, one can similarly adopt equalized odds as the measure of fairness. Equalized odds as the measure of fairness is defined as:

$$\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}) = 1, s = i |\, y = k) = \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}) = 1, s = j |\, y = k) \quad \forall\, i, j \in \mathcal{S},\ k \in \mathcal{Y} \tag{13}$$

Therefore, we must add a regularizer per each class label to satisfy the equalized odds notion. Other notions of fairness can be used in this framework as long as they can be represented as the conditional independence between sensitive attributes, predictions, and labels [9].

Appendix C. f-divergences for Continuous Sensitive Attributes and Target Variables

In Section 2, we developed a framework for promoting fairness for classification problems where both target labels and sensitive attributes are discrete variables. Hence we could efficiently solve the variational formulation that arose through the designing of unbiased estimators. However, it is not uncommon to find applications of f-divergence regularizers in practice that require either the sensitive features or the output variable to be continuous; or both to be continuous parameters. In such cases, the summation over the respective variable is replaced by an integral over the probability distribution. The challenging aspect is calculating the variational form's integral and trailing supremum in the continuous domain.

Let P and Q be two continuous distributions over the space Ω such that P is absolutely continuous with respect to Q ($P \ll Q$). Then, the f-divergence between these two distributions for a given convex function f is defined as:

$$\mathcal{D}_f(\mathbb{P}, \mathbb{Q}) = \int_{\Omega} f\Big(\frac{dP}{dQ}\Big) dQ \tag{14}$$

When the target variable is continuous (regression problems), but the sensitive attribute is discrete, (f-FERM) can be written as:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A} \in \mathbb{R}^{\infty}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_{k} \int_{x} \Big[\mathbf{A}_k(x)\mathbb{P}(x) - f^*(\mathbf{A}_k(x))\mathbb{Q}_k\Big] dx$$

With slight changes, the above problem can be reformulated as follows:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A} \in \mathbb{R}^{jk}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \max_{\mathbf{A}_1, \ldots, \mathbf{A}_m} \quad \sum_{k} \mathbb{E}\Big[\mathbf{A}_k(s)\mathbb{P}_j(s) - f^*(\mathbf{A}_k(s))\mathbb{Q}_k\Big]$$

When both sensitive features and target variables are continuous, the objective function becomes:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A} \in \mathbb{R}^{\infty \times \infty}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \int_{x} \int_{y} \Big[\mathbf{A}(x, y)\mathbb{P}(x) - f^*(\mathbf{A}(x, y))\mathbb{Q}(y)\Big] dx\ dy$$

Such a formulation is clearly intractable for solving $\mathbf{A}_k(x)$ or $\mathbf{A}(x, y)$ in the continuous domain. We need to approximate the above integrals in discretized/quantized regions or find another variational representation for designing unbiased estimators of continuous domain f-divergences. We leave developing algorithms for the continuous target variables and sensitive attributes as a future direction.

Appendix D. f-divergences Cover Well-known Notions of Fairness Violation

In this section, we show that optimizing f-divergences to 0 guarantees the independence of the sensitive attributes and predictions. In other words, optimizing f-divergences leads to a fair model under the demographic parity notion (or other group fairness notions discussed in Appendix B).

Proposition 4 *[43, Theorem 2.3] Let f be a convex function from $\mathbb{R}^+$ to $\mathbb{R}$, such that f is convex, $f(1) = 0$, f is strictly convex in a neighborhood of 1. Then $\mathcal{D}_f(\mathbb{P}||\mathbb{Q}) = 0$, if and only if $P = Q$.*

As an immediate result, a trained model in (f-FERM) is fair under the demographic notion if and only if

$$\mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s) = \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x})) \otimes \mathbb{P}(s), \tag{15}$$

which means the independence of s and $\hat{y}_{\boldsymbol{\theta}}(\mathbf{x})$.

Next, we show f-divergences either include or provide an upper bound for well-known notions of fairness violation in the literature.

Proposition 5 *Exponential Rényi Mutual Information (ERMI) [35, 38] is an f-divergence with $f(t) = (t-1)^2$*

Proof Exponential Rényi Mutual Information is defined as [35]:

$$\text{ERMI}(\hat{y}, s) = \sum_{j \in \mathcal{Y}, k \in \mathcal{S}} \frac{\hat{P}_{\hat{y},s}(j,k)^2}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)} - 1 \tag{16}$$

For the case of $f(t) = (t-1)^2$, we have:

$$\begin{aligned}
\mathcal{D}_f\big(\hat{P}_{\hat{y}} \otimes \hat{P}_s || \hat{P}_{\hat{y},s}\big) &= \sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \hat{P}_{\hat{y}}(j)\hat{P}_s(k) f\Big(\frac{\hat{P}_{\hat{y},s}(j,k)}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)}\Big) = \sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \hat{P}_{\hat{y}}(j)\hat{P}_s(k)\Big(\frac{\hat{P}_{\hat{y},s}(j,k)}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)} - 1\Big)^2 \\
&= \sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \hat{P}_{\hat{y}}(j)\hat{P}_s(k)\Big(\frac{\hat{P}_{\hat{y},s}(j,k)^2}{\hat{P}_{\hat{y}}(j)^2\hat{P}_s(k)^2} - 2\frac{\hat{P}_{\hat{y},s}(j,k)}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)} + 1\Big) \\
&= \sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \Big(\frac{\hat{P}_{\hat{y},s}(j,k)^2}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)} - 2\hat{P}_{\hat{y},s}(j,k) + \hat{P}_{\hat{y}}(j)\hat{P}_s(k)\Big) \\
&= \sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \frac{\hat{P}_{\hat{y},s}(j,k)^2}{\hat{P}_{\hat{y}}(j)\hat{P}_s(k)} - 2 + 1 = \text{ERMI}(\hat{y}, s)
\end{aligned}$$

Note that, in the last equality, we use:

$$\sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \hat{P}_{\hat{y},s}(j,k) = \sum_{j \in \mathcal{Y}} \hat{P}_{\hat{y}}(j) = 1,$$

and

$$\sum_{j \in \mathcal{Y}} \sum_{k \in \mathcal{S}} \hat{P}_{\hat{y}}(j)\hat{P}_s(k) = \sum_{j \in \mathcal{Y}} \hat{P}_{\hat{y}}(j)\Big(\sum_{k \in \mathcal{S}} \hat{P}_s(k)\Big) = \sum_{j \in \mathcal{Y}} \hat{P}_{\hat{y}}(j) = 1,$$

■

Proposition 6 *Demographic parity violation is upper-bounded by the f-divergence for $f(t) = (t-1)^2$*

Proof Based on Proposition 5, ERMI is an f-divergence with $f(t) = (t-1)^2$. Therefore, the proposition is an immediate result of Lemma 3 in [35]. ■

Proposition 7 *Rényi correlation [5] can be upper bounded by the f-divergences for the choice of $f(t) = (t-1)^2$.*

Proof Based on Proposition 5, ERMI is an f-divergence with $f(t) = (t-1)^2$. Therefore, the proposition is an immediate result of Lemma 2 in [35]. ■

Remark 8 *Mutual Information as the measure of fairness violation [11] is a special case of f-divergences for the choice of KL-divergence $f(t) = t\log(t)$ in (f-FERM).*

Appendix E. Proof of Proposition 1

Lemma 9 *Assume that $f(\mathbf{z})$ is a semi-continuous convex function. Therefore, f can be written as the following maximization problem:*

$$f(\mathbf{z}) = \max_{\alpha} \mathbf{z}^T\alpha - g(\alpha)$$

where g is the convex conjugate of f.

Proof Let g be the convex conjugate of the function f defined as:

$$g(\alpha) = \sup_{\mathbf{z}} \boldsymbol{\alpha}^T\mathbf{z} - f(\mathbf{z})$$

Since f is a lower semi-continuous convex function, by Fenchel-Moreau theorem [25], it is biconjugate, which means the taking conjugate of g transforms it back to f. Therefore,

$$f(\mathbf{z}) = \sup_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^T\mathbf{z} - g(\boldsymbol{\alpha})$$

where g is the convex conjugate of f. ■

Based on the above lemma, we have:

$$\mathcal{D}_f(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{m}\mathbb{Q}_i f\Big(\frac{\mathbb{P}_i}{\mathbb{Q}_i}\Big) = \mathcal{D}_f(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{m}\mathbb{Q}_i \sup_{\alpha_i\in\text{dom}f} \alpha_i\frac{\mathbb{P}_i}{\mathbb{Q}_i} - f^*(\alpha_i)$$
$$= \sup_{\alpha_1,\ldots,\alpha_m\in\text{dom}f} \sum_{i=1}^{m}\alpha_i\mathbb{P}_i - f^*(\alpha_i)\mathbb{Q}_i$$

Set $\mathbb{P} = \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), s), \mathbb{Q} = \mathbb{P}(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}))\otimes\mathbb{P}(s)$, and $\alpha_i = \mathbf{A}_{jk}$. Therefore, we obtain the formulation in (3).

Appendix F. Derivation of Closed-Form Expressions for Unbiased Gradient Estimators of f-Divergences

Proposition 10 *For two functions $f(t), g(t)$ such that $g(t) = f(t) + c(t-1)$, then $\mathcal{D}_f(\cdot|\cdot) \equiv \mathcal{D}_g(\cdot|\cdot)$.*

Proof Proof follows naturally from [43, Proposition 7.2] ■

Theorem 11 *Let $f(t) = (t-1)^2$ and $\mathbb{P}(s=k) = \pi_k$ (χ^2 Divergence). Then, Equation (1) can be written as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[A_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) - (\mathbf{A}_{jk} + \frac{\mathbf{A}_{jk}^2}{4}) \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big] \tag{17}$$

Variational Representation of $f(x) = (x-1)^2$ is given by

$$f(x) = \sup_{\alpha} (\alpha x - f^*(\alpha))$$

Where $f^*(\alpha)$ is the convex conjugate

$$f^*(\alpha) = \sup_{x} (x\alpha - f(x))$$

Taking derivative of $f^*(\alpha)$ w.r.t x gives $x^* = \alpha/2 + 1$. This results in $f^*(\alpha) = \alpha + \alpha^2/4$

Theorem 12 *Let $f(t) = -\ln(t)$ and $\mathbb{P}(s=k) = \pi_k$ (Reverse KL). Then, Equation (1) can be written as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[\mathbf{A}_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) + (1 + \ln(-\mathbf{A}_{jk})) \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big] \tag{18}$$

Proceeding as above, optimal x^* for the supremum of $f^*(\alpha)$ is $x^* = -1/\alpha$,
resulting in $f^*(\alpha) = -1 - \ln(-\alpha)$.

Theorem 13 *Let $f(t) = \frac{1}{2}|t-1|$ and $\mathbb{P}(s=k) = \pi_k$ (Total Variational Distance). Then, Equation (1) can be written as (where $|\mathbf{A}_{jk}| \leq \frac{1}{2}$):*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \mathbf{A}_{jk} \Big[\mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) - \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big] \tag{19}$$

For $f = \frac{1}{2}|t-1|$, the variational representation is $f(x) = \sup_\alpha \Big(\alpha x - f^*(\alpha) \Big)$
Through the convex conjugate $f^*(\alpha)$, we have that

$$\begin{aligned} f^*(\alpha) &= \sup_x \Big(x\alpha - f(\alpha) \Big) = \sup_x \Big(x\alpha - \frac{1}{2}|x-1| \Big) \\ &= \begin{cases} \infty & \text{for } |\alpha| > \frac{1}{2} \\ \alpha & \text{for } |\alpha| \leq \frac{1}{2} \end{cases} \end{aligned}$$

So $|\alpha| \leq \frac{1}{2}$ is constrained for the supremum/maximum to exist (otherwise tends to ∞).

Theorem 14 *Let $f(t) = t\ln(t)$ and $\mathbb{P}(s = k) = \pi_k$ (KL Divergence). Then, Equation (1) can be written as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[\mathbf{A}_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) - e^{\mathbf{A}_{jk}-1} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big] \tag{20}$$

For $f(t) = t\ln(t)$ in f-divergence, the convex conjugate can be represented by:

$$f^*(\alpha) = \sup_x (x\alpha - x\ln(x))$$

On differenting w.r.t x for attaining supremum, we get $x = e^{\alpha-1}$. Hence, the variational representation of $f(t) = t\ln(t)$ becomes:

$$f(x) = \sup_{\alpha} \left(x\alpha - e^{\alpha-1} \right)$$

Note: We can also use the affine transformation $\alpha \leftarrow \alpha - 1$ which results in the more commonly studied version in literature:

$$D(P||Q) = 1 + \sup_{g:X\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[e^{g(X)}]$$

Theorem 15 *Let $f(t) = -(t+1)\ln(\frac{t+1}{2}) + t\ln(t)$ and $\mathbb{P}(s = k) = \pi_k$ (Jensen-Shannon Divergence). Then, Equation (1) can be written as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[\mathbf{A}_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) + \ln(2 - e^{\mathbf{A}_{jk}}) \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big] \tag{21}$$

For the JS Divergence, we have $f(t) = -(t+1)\ln(\frac{t+1}{2}) + t\ln(t)$, whose convex conjugate can be represented as:

$$f^*(\alpha) = \sup_x \left(\alpha x + (x+1)\ln\left(\frac{x+1}{2}\right) - x\ln(x) \right)$$

On differentiating w.r.t x to obtain the supremum, we have

$$\frac{2x}{x+1} = e^{\alpha} \implies x = \frac{e^{\alpha}}{2 - e^{\alpha}}$$

Substituting x in $f^*(\alpha)$,

$$f^*(\alpha) = -\ln(2 - e^{\alpha})$$

Thus, in $f(x) = \sup_{\alpha} \left(x\alpha - f^*(\alpha) \right)$, we get the variational form as:

$$f(x) = \sup_{\alpha} \left(x\alpha + \ln(2 - e^{\alpha}) \right)$$

Theorem 16 *Let $f(t)$ be*

$$f(t) = \begin{cases} \frac{t^\alpha - \alpha t - (1-\alpha)}{\alpha(\alpha-1)} & \text{if } \alpha \neq 0, \alpha \neq 1 \\ t\ln(t) - t + 1 & \text{if } \alpha = 1 \\ -\ln(t) + t - 1 & \text{if } \alpha = 0 \end{cases}$$

and $\mathbb{P}(s = k) = \pi_k$ (General α Divergence). Then, Equation (1) can be written as:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[\mathbf{A}_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) - \frac{\mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j)}{\alpha} \Big(\Big((\alpha-1)\mathbf{A}_{jk} + 1 \Big)^{\frac{\alpha}{\alpha-1}} - 1 \Big) \Big] \quad (22)$$

Excluding the limiting cases where $\alpha = 1$ or $\alpha = 0$, we can find the convex conjugate $f^*(y)$ as:

$$\begin{aligned} f^*(y) &= \sup_x \Big(xy - f(x) \Big) \\ &= \sup_x \Big(xy - \frac{x^\alpha - \alpha x - (1-\alpha)}{\alpha(\alpha-1)} \Big) \end{aligned}$$

On differentiating w.r.t. x, we obtain (here variational parameter is y, do not confuse with the constant α)

$$x^* = \Big((\alpha-1)y + 1 \Big)^{\frac{1}{\alpha-1}}$$

Thus,

$$f^*(y) = \frac{\Big((\alpha-1)y + 1 \Big)^{\frac{\alpha}{\alpha-1}}}{\alpha} - \frac{1}{\alpha}$$

KL Divergence and Reverse KL Divergence can be obtained by taking the limit when α tends to 1 and 0, respectively.

Note: Standard literature on divergences often parametrize the α-divergence as

$$f(x) = \begin{cases} t\ln(t) & \text{if } \alpha = 1 \\ -\ln(t) & \text{if } \alpha = -1 \\ \frac{4}{1-\alpha^2}\Big(1 - t^{(1+\alpha/2)}\Big) & \text{otherwise} \end{cases}$$

This is tantamount to the substitution $\alpha \leftarrow \frac{1+\alpha}{2}$ in the original definition of generalized f-divergence.

Theorem 17 *Let $f(t) = (\sqrt{t} - 1)^2$ (equivalently $f(t) = 2(1 - \sqrt{t})$) and $\mathbb{P}(s = k) = \pi_k$ (Squared Hellinger Distance). Then, Equation (1) can be written as:*

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{A}} \sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \sum_j \sum_k \pi_k \Big[\mathbf{A}_{jk} \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j | s = k) + \mathbb{P}(\hat{y}_{\boldsymbol{\theta}} = j) \Big(\frac{1}{\mathbf{A}_{jk}} + 2 \Big) \Big] \quad (23)$$

For Squared Hellinger Distance,

$$\begin{aligned} f^*(\alpha) &= \sup_x (x\alpha - f(x)) \\ &= \sup_x (x\alpha - 2(1-\sqrt{x})) \end{aligned}$$

On differentiating w.r.t. x, we get

$$\alpha + \frac{1}{\sqrt{x}} = 0 \text{ (Note } \alpha < 0) \implies x = \frac{1}{(\alpha)^2}$$
$$\implies f^*(\alpha) = \frac{\alpha}{\alpha^2} - 2 + \frac{(-2)}{\alpha} = \frac{-1}{\alpha} - 2$$

Note that the first, second, and third terms are negative, negative, and positive, respectively; hence the appropriate choice of sign(α) for functions of odd powers of α.

Appendix G. Formal Statement of Theorem 2 and Proof

Theorem 18 ***Formal Statement of Theorem*** *Let* $(\mathbf{x}_i, y_i, s_i) \quad \forall 1 \leq i \leq n$ *be the collection of* n *data points satisfying the following assumptions:*

- $\ell(\cdot, \mathbf{x}, y)$ *is* G*-Lipschitz, and* β_ℓ*-smooth for all* $\mathbf{x}_i, y_i$*.*
- $F_j(\cdot, \boldsymbol{\theta})$ *is* L*-Lipschitz and* b*-smooth for all* $\boldsymbol{\theta}$ *and all label classes* j*.*
- $\widehat{p}_{\hat{y}}^{\min} := \inf_{\{\boldsymbol{\theta}^t, t\in[T]\}} \min_{j\in[m]} \frac{1}{N}\sum_{i=1}^N \hat{y}_{\boldsymbol{\theta},j}(\mathbf{x}_i) \geq \frac{\mu}{2} > 0.$
- $\hat{p}_S^{\min} := \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{s_i=j\}} > 0.$

choose $\eta_{\boldsymbol{\theta}} = \Theta(\frac{\epsilon^4}{\ell^3 L^2 D^2})$ *and* $\eta_\alpha = \Theta(\frac{\epsilon^2}{\ell\sigma^2})$ *and the mini-batch size of* 1*. Therefore, Algorithm 1 finds an* ϵ*-stationary of Problem* f*-FERM in* $\mathcal{O}(\frac{1}{\epsilon^8})$*.*

Remark 19 *The* ***first assumption*** *listed in the theorem statement is true for popular losses such as cross-entropy loss and squared loss (assuming that the input data takes values in a bounded set, which holds for all real-world datasets).*

Remark 20 *The* ***second assumption*** *holds for popular classifiers generating probability vectors (e.g., logits in neural networks, logistic regression outputs). For classifiers with no probability output, one must transform the output to a number between zero and one first.*

Remark 21 *The* ***third assumption*** *states that the probability of assigning a label to the data points must not be zero for all data points for any label at each iteration.*

Remark 22 *Finally, the* ***fourth assumption*** *ensures each sensitive class's probability is not zero. In other words, there should be at least one point in the dataset with that sensitive attribute for any sensitive group. It holds for all benchmark datasets in practice. Simply put, any protected group appearing during the test phase must have at least one representative in the training data.*

The following lemma is helpful for the proof of the theorem:

Lemma 23 *Let $A_1, \ldots, A_n$ be n variables such that $\|A_i\|_2 \leq c_i$. Then, we have:*

$$\mathbb{E}[\|\sum_{i=1}^{n} A_i\|_2^2] \leq n \sum_{i=1}^{n} c_i^2 \tag{24}$$

Proof

$$\|\sum_{i=1}^{n} A_i\|_2^2 = \sum \|A_i\|_2^2 + 2\sum_{i\neq j}\langle A_i, A_j\rangle \leq \sum \|A_i\|_2^2 + \sum_{i\neq j} \|A_i\|_2^2 + \|A_j\|_2^2 = n\sum_{i=1}^{n}\|A_i\|_2^2,$$

which is based on the fact that $2\langle A_i, A_j\rangle \leq \|A_i\|_2^2 + \|A_j\|_2^2$. Therefore:

$$\mathbb{E}[\|\sum_{i=1}^{n} A_i\|_2^2] \leq n\sum_{i=1}^{n}\mathbb{E}[\|A_i\|_2^2] \leq n\sum_{i=1}^{n} c_i^2$$

■

Now, we are ready to prove Theorem 18.

Proof The proof consists of three main steps. First, we need to show that the gradient estimator in Algorithm 1 is unbiased. Since the samples are IID, for any function $\psi(\cdot,\cdot)$, and an IID batch of data points $\mathcal{B}$ we have:

$$\mathbb{E}\Big[\frac{1}{\mathcal{B}}\sum_{(\mathbf{x},y)\in\mathcal{B}} \nabla\psi(\mathbf{x},y)\Big] = \frac{1}{\mathcal{B}}\sum_{(\mathbf{x},y)}\mathbb{E}[\psi(\mathbf{x},y)] = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}(\mathbf{x},y,s)}[\nabla\psi(\mathbf{x},y)]$$

As an immediate result, if the objective function is written as the summation over n functions, the gradient estimator over an IID batch of data will be unbiased. According to Equation (4), the objective function has the desired form for:

$$\min_{\boldsymbol{\theta}}\max_{\mathbf{A}} \quad \frac{1}{n}\sum_{i=1}^{n}\left[\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda\sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}}\Big[\mathbf{A}_{jk}F_j(\mathbf{x}_i;\boldsymbol{\theta})\mathbb{1}(s_i=k) - f^*(\mathbf{A}_{jk})\pi_k F_j(\mathbf{x}_i;\boldsymbol{\theta})\Big]\right] \tag{25}$$

Next, we need to show the boundedness of the gradient estimator variance. Let

$$G_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{(x_i,y_i)\in\mathcal{B}}\left[\nabla_{\boldsymbol{\theta}}\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda\sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}}\Big[\mathbf{A}_{jk}\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\mathbb{1}(s_i=k) - f^*(\mathbf{A}_{jk})\pi_k\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\right]\right]$$

We need to show for a given data batch:

$$\mathbb{E}[\|G_{\mathcal{B}} - G_n\|_2^2]$$

where G_n is the gradient with respect to all n data points (when $\mathcal{B} = \{1,\ldots,n\}$. Note that:

$$\|G_{\mathcal{B}} - G_n\|_2^2 \leq 2\|G_{\mathcal{B}}\|_2^2 + \|G_n\|_2^2$$

Thus, it suffices to show that the gradient is bounded for any given $\mathcal{B}$ batch. Since the samples are independent of each other and identically distributed from $\mathbb{P}_{\text{train}}$ (IID samples), the second-order moment of the average over $|\mathcal{B}|$ data points is $1/|\mathcal{B}|$ times the variance of a single data point.

Thus, we need to show that the boundedness of the gradient for a given data point drawn from the training distribution:

$$\left[\nabla_{\boldsymbol{\theta}}\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), y_i) + \lambda \sum_{\substack{j\in\mathcal{Y},\\ k\in\mathcal{S}}} \left[\mathbf{A}_{jk}\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\mathbb{1}(s_i=k) - f^*(\mathbf{A}_{jk})\pi_k\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\right]\right] \tag{26}$$

Based on the first assumption:

$$\|\nabla_{\boldsymbol{\theta}}\ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}), y_i)\|_2 \leq G \tag{27}$$

Based on the second assumption:

$$\|\mathbf{A}_{jk}\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\mathbb{1}(s_i=k)\|_2 \leq L\mathbf{A}_{jk} \tag{28}$$

$$\|\pi_k f^*(\mathbf{A}_{jk})\nabla_{\boldsymbol{\theta}}F_j(\mathbf{x}_i;\boldsymbol{\theta})\|_2 \leq \pi_k L f^*(\mathbf{A}_{jk}) \tag{29}$$

These terms are bounded if $\mathbf{A}_{jk}$ is bounded and $f^*(\mathbf{A}_{jk})$ is bounded for any $\mathbf{A}_{jk}$. This holds true for all f-divergences given assumptions 3 and 4. To see why, it suffices to find the optimal solution of each f-divergence by setting the gradient zero with respect to $\mathbf{A}_{jk}$. In all cases, the solution is a combination of $\mathbb{P}_{s_k}$ and $\mathbb{P}_{\hat{y}_j}$ terms that are non-zero and bounded (by assumptions 3 and 4). Since each term is bounded in (26), the expectation of the squared norm is also bounded, according to Lemma 23.

Finally, given that the estimator is unbiased, and the variance is bounded (Assumption 4.1 holds in Lin et al. [34]), the two-time-scale stochastic gradient descent-ascent algorithm (which is Algorithm 1) finds an ϵ-stationary solution of the Problem in $\mathcal{O}(\frac{1}{\epsilon^8})$ according to Theorem 4.9 in Lin et al. [34]. ■

Appendix H. Robust f-FERM in the Presence of Distribution Shifts

A semi-stochastic memory-efficient first-order training algorithm. To apply (stochastic) gradient descent-ascent algorithm [34] to problem (8), we need to have unbiased estimator of the function $h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V})$ w.r.t. $\boldsymbol{\theta}$, $\mathbb{U}$, and $\mathbb{V}$ variables. While it seems challenging to obtain unbiased estimator w.r.t. all variables, one can notice that if $\hat{p}_j(\boldsymbol{\theta})$ and $\hat{q}_j(\boldsymbol{\theta})$ can be computed easily with one forward pass over all data points (i.e., in $O(m \times n)$ memory requirement). Consequently, the gradient of $h(\boldsymbol{\theta}, \mathbb{U}, \mathbb{V})$ w.r.t. $\mathbb{U}$ and $\mathbb{V}$ can be computed with one forward pass over all data points (without the need for doing backpropagation). On the other hand, one can easily obtain unbiased estimator of $\nabla_{\boldsymbol{\theta}}\hat{p}_j(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}}\hat{q}_j(\boldsymbol{\theta})$ in (10) using a small mini-batch of data. Such a task requires $O(b \times d)$ memory with d being the number of parameters (i.e., $\boldsymbol{\theta} \in \mathbb{R}^d$) and b being the batch size. Combining this unbiased estimation with the computed values of $\hat{p}_j(\boldsymbol{\theta})$ and $\hat{q}_j(\boldsymbol{\theta})$ leads to an unbiased estimator of the objective of (8) w.r.t. $\boldsymbol{\theta}$ variable. To summarize, we need to do one forward propagation to obtain gradients w.r.t. $\mathbb{U}$ and $\mathbb{V}$, and we only do backpropagation for computing gradients w.r.t. $\boldsymbol{\theta}$ over the mini-batch of data. Such an algorithm requires $O(mn + bd)$ memory requirement; and thus can be used for training large models (with $d, n \gg b, m$). It is known that memory requirements are the major limiting factors in training large models such as LLMs [37].

H.1. Robust f-FERM Under ℓ_∞ Norms and Potentially Large Distribution Shifts

The developed framework in the previous section assumes the distribution shift is small (the uncertainty set diameter is smaller than a certain threshold). When preserving fairness in the presence of large distribution shifts is a priority, our previous methodology might not work well. As discussed before, the formulation (7) leads to a nonconvex-nonconcave min-max optimization problem and this class of problems is hard to solve computationally in general (even to stationarity notions). Thus, we need to exploit the structure of the problem. In this section, we show that we can exploit the structure to develop a first-order algorithm under large distribution shifts. Particularly, we focus on the case where the uncertainty set is ℓ_∞ ball and the divergence satisfies certain assumptions (i.e., $f^*(\alpha^*) > 0$ and $\alpha^* > 0$, which is satisfied for KL divergence).

For the general f-divergence in (7), it is easy to show that the function D_f is convex in $\mathbb{P}$ and $\mathbb{Q}$. Thus, under ℓ_∞ uncertainty set on $\mathbb{P}$ and $\mathbb{Q}$, the optimal solution of the maximization problem in (7) will be at an extreme point. Moreover, under the assumption that $f^*(\alpha^*) > 0$ and $\alpha^* > 0$ (which is satisfied for KL divergence), one can easily see that the optimal $p_j = \min\{\hat{p}_j + \delta, 1\}$ and $q_j = \max\{\hat{q}_j - \delta, 0\}$. Notice that to obtain this efficient optimal closed-form solution, we need to relax the probability simplex constraint. Thus under this assumption, problem (7) can be reformulated as

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \lambda \mathcal{D}_f(\min\{\mathbb{P}+\delta, 1\} || \max\{\mathbb{Q}-\delta, 0\}), \tag{30}$$

which is a regular minimization problem and (sub)gradient descent can be utilized to solve it.

Algorithm 2 Gradient-Regularization Robust Training algorithm

Input: $\boldsymbol{\theta}^0 \in \mathbb{R}^{d_\theta}$, step-sizes $\eta_{\boldsymbol{\theta}}$, η_α, fairness parameter $\lambda \geq 0$, iteration number T, Batchsize $[b]_t$
for $t = 1, \ldots, T$ **do**
  Sample minibatch of data $\mathbf{b}_t = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_b, \mathbf{y}_b)\}$
  Estimate $\mathbb{P}(\hat{\mathbf{y}}_{\boldsymbol{\theta}^t})$ for minibatch $\mathbf{b}_t$
  repeat
    $d\mathbf{A}_{jk} = \nabla_{\mathbf{A}}(\mathbf{A}_{jk}\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}},\mathbf{s}} - f^*(\mathbf{A}_{jk})\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}\mathbb{P}_{\mathbf{s}})$
    $\mathbf{A}_{jk} = \mathbf{A}_{jk} + \eta_\alpha \, d\mathbf{A}_{jk}$
  until Convergence to $\mathbf{A}^*_{jk}$
  Obtain closed form expressions: $\frac{\partial}{\partial \boldsymbol{\theta}} ||\nabla_{\mathbb{P}} \mathcal{D}_f(\mathbb{P}||\mathbb{Q})||_2^2$ and $\frac{\partial}{\partial \boldsymbol{\theta}} ||\nabla_{\mathbb{P}} \mathcal{D}_f(\mathbb{P}||\mathbb{Q})||_2^2$ in terms of $\mathbb{P}_{\hat{\mathbf{y}}_{\boldsymbol{\theta}}}$
  $d\boldsymbol{\theta} = \nabla_{\boldsymbol{\theta}}\Big[\ell(\boldsymbol{\theta}^{t-1}, \mathbf{x}, \mathbf{y}) + \lambda\Big[\mathcal{D}_f(\hat{\mathbb{P}}||\hat{\mathbb{Q}}) + \epsilon\Big(||\nabla_{\mathbb{P}} \mathcal{D}_f(\hat{\mathbb{P}}||\hat{\mathbb{Q}})||_2^2 + ||\nabla_{\mathbb{Q}} \mathcal{D}_f(\hat{\mathbb{P}}||\hat{\mathbb{Q}})||_2^2\Big)\Big]\Big]$
  $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta_{\boldsymbol{\theta}} \, d\boldsymbol{\theta}$
Return: $\boldsymbol{\theta}^T$

H.2. Fairness-Accuracy Tradeoffs in the Presence of the Distribution Shift

We consider two experiments to evaluate the robustness of Algorithms developed above. In the first experiment, we randomly switch the label of genders for $n\%$ of the data points (n ranges from 1 to 20) in the Adult dataset. Then, we train models on the new datasets with a proportion of corrupted sensitive attributes and evaluate the performance on the test data. Figure 3 is obtained by training different models to achieve 80% accuracy on the test data and comparing their demographic parity violation. By increasing the percentage of corrupted sensitive attributes, we see that both

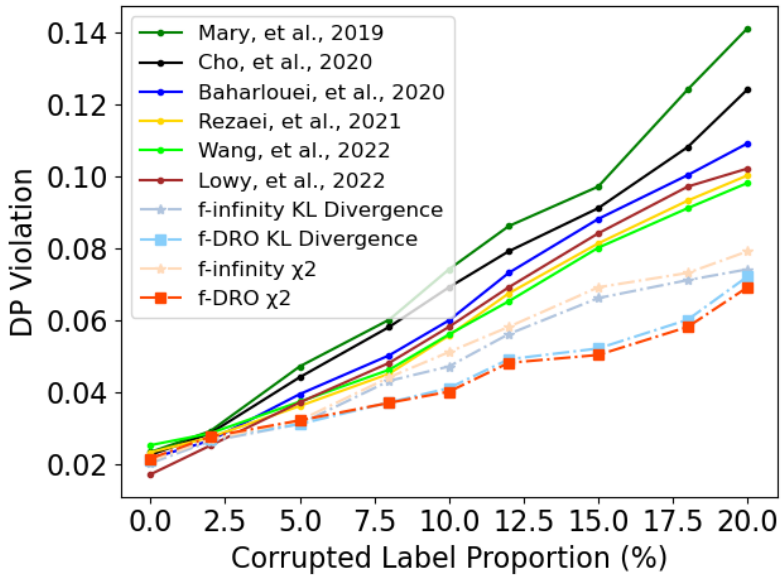

Figure 3: Performance of different state-of-the-art approaches and our two methods for handling distribution shift. The dataset is adult, and the sensitive attribute is gender. We randomly flip the label of a proportion of gender entries (from 0 to 20%). As we observe, our approach demonstrates more robustness against the drop in DP violation compared to other approaches.

f-DRO and f-infinity achieve less DP violation than SOTA approaches in the literature. In this specific experiment, f-DRO works better than f-infinity, and there is no significant difference between choosing KL-divergence or χ^2 as the function f. Among the papers designed for handling distribution shifts, Rezaei et al. [48] and Wang et al. [59] were the only options with the available implementation. In a more recently collected dataset (new adult) [15], the users are separated based on their living state. We train different fair models in a single state and evaluate the fairness-accuracy tradeoff in other states. Figure 4 depicts the performance of different methods. For each method, the center point is the average of accuracy and fairness among 50 states. The horizontal and vertical lines show the 25-percentile to 75-percentile range of performance among the states. The training fairness violation is set to 0.02 for all methods. We observe that f-infinity preserves the fairness level better than other approaches. In comparison, f-DRO has a better accuracy. Depending on the application, we suggest using f-infinity if preserving a high level of fairness is a priority and f-DRO for the cases when a better tradeoff between fairness and accuracy is expected. Note that both of these approaches offer better fairness-accuracy tradeoffs compared to the SOTA approaches in the literature.

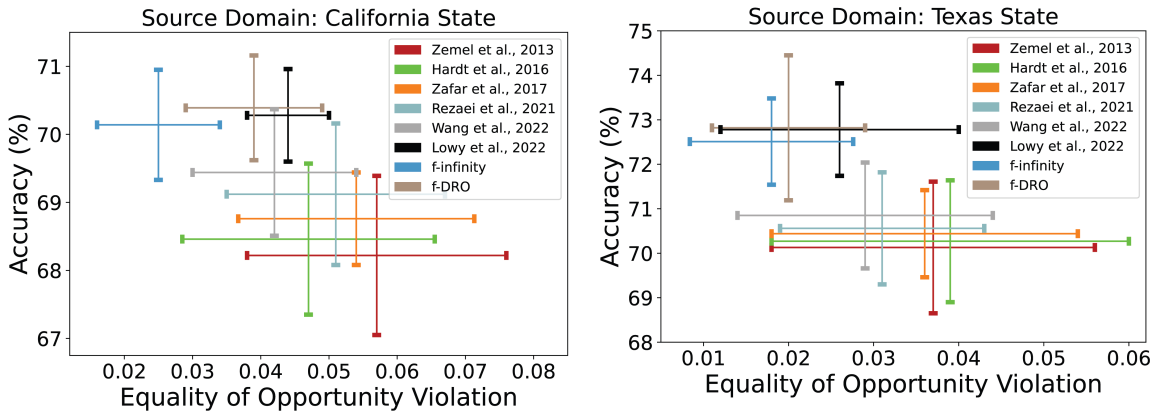

Figure 4: Performance of the trained fair models on new Adult Dataset. The model is trained on one state (California or Texas) and evaluated in 50 states. The distribution of each state dataset is different than others. Thus, the IID assumption does not hold among datasets of different states.

Appendix I. Details of Tuning Hyperparameters

In all experiments, we set $\eta_{\boldsymbol{\theta}} = 10^{-5}$ and $\eta_{\alpha} = 10^{-6}$. Further, we train the model with $\lambda = 0$ for 300 epochs, and then we set λ to the considered value. We continue the training until 2000 epochs. The range of λ to get each point in the tradeoff figures is varied for different f-divergences. The KL-divergence λ range is $[0, 150]$. For χ^2 divergence it is $[0, 300]$ and for the reverse KL it is $[0, 50]$. Moreover, the λ range for JS and Squared Hellinger is $[0, 110]$ and $[0, 250]$. Note that larger values outside the range lead to models with 0 predictions for all values.

In the DRO case, aside from λ we must tune ϵ, the robustness parameter. To achieve the best result, we have two different strategies depending on the availability of the data from the target domain. Suppose we have access to a collection of data points from the target domain. In that case, we consider it as the validation set to choose the optimal combination of $\lambda \in \{0.1, 0.5, 1, 2, 5, 10, 20, 50\}$ and $\delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$. In the second scenario, when we do not have any access to target domain data, we perform a k-fold cross-validation on the source data. A more elegant way is to create the validation dataset by oversampling the minority groups. Having access to the oversampled validation set, we choose the optimal λ and δ similar to the first scenario. In the experiment regarding Figure 4, we reserve 5% of data from the target domain for validation (scenario 1). In Figure 2, we apply scenario 2 to tune the hyperparameters λ and δ.

Appendix J. Further Experiments on Other Datasets and Notions of Fairness

In this section, we perform (f-FERM), [23], [38], and [5] to COMPAS [2] and German Credit datasets [3]. In the experiment on COMPAS, we use equality of opportunity as the measure of fairness violation, while in the German Credit dataset experiment, we use equalized odds. The results show that (f-FERM) is significantly better than other approaches regarding the accuracy-fairness tradeoff. The batch size is equal to 64 for all methods.

2. https://www.kaggle.com/datasets/danofer/compass
3. https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

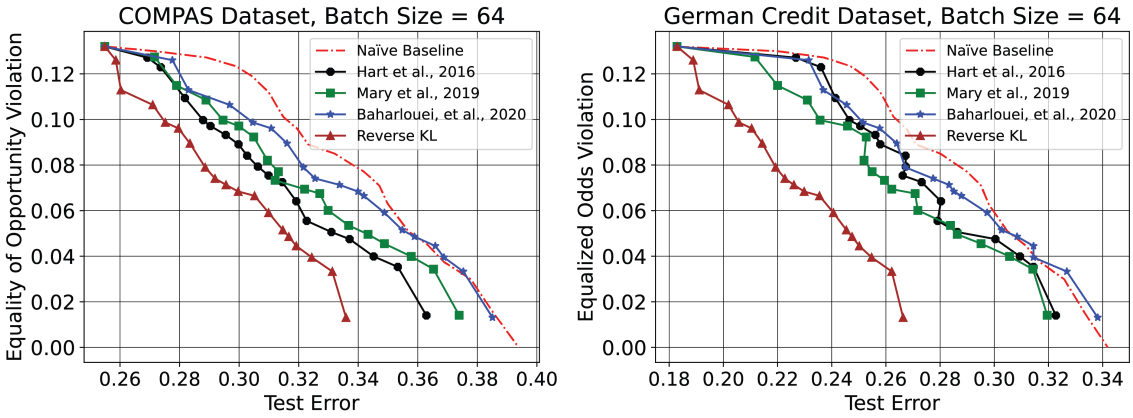

Figure 5: Performance of the trained fair models on COMPAS and German Credit Datasets.