# FERMI: Fair Empirical Risk Minimization Via Exponential Rényi Mutual Information

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Despite the success of large-scale empirical risk minimization (ERM) at achieving high accuracy across a variety of machine learning tasks, fair ERM is hindered by the incompatibility of fairness constraints with stochastic optimization. In this paper, we propose the fair empirical risk minimization via exponential Rényi mutual information (FERMI) framework. FERMI is built on a stochastic estimator for exponential Rényi mutual information (ERMI), an information divergence measuring the degree of the dependence of predictions on sensitive attributes. Theoretically, we show that ERMI upper bounds existing popular fairness violation metrics, thus controlling ERMI provides guarantees on other commonly used violations, such as $L_\infty$. We derive an unbiased estimator for ERMI, which we use to derive the FERMI algorithm. We prove that FERMI converges for demographic parity, equalized odds, and equal opportunity notions of fairness in stochastic optimization. Empirically, we show that FERMI is amenable to large-scale problems with multiple (non-binary) sensitive attributes and non-binary targets. Extensive experiments show that FERMI achieves the most favorable tradeoffs between fairness violation and test accuracy across all tested setups compared with state-of-the-art baselines for demographic parity, equalized odds, equal opportunity. These benefits are especially significant for non-binary classification with large sensitive sets and small batch sizes, showcasing the effectiveness of the FERMI objective and the developed stochastic algorithm for solving it.

## 1 Introduction

Ensuring that decisions made using machine learning algorithms are fair to different subgroups is of utmost importance. Without any mitigation strategy, machine learning algorithms may result in discrimination against certain subgroups based on sensitive attributes, such as gender or race, even if such discrimination is absent in the training data (Datta et al., 2015; Sweeney, 2013; Bolukbasi et al., 2016; Angwin et al., 2016; Calmon et al., 2017b; Feldman et al., 2015; Hardt et al., 2016; Fish et al., 2016; Woodworth et al., 2017; Zafar et al., 2017; Bechavod & Ligett, 2017; Kearns et al., 2018). Algorithmic fairness literature aims to remedy such discrimination issues.

A machine learning algorithm satisfies the *demographic parity* fairness notion, if the predicted target is independent of the sensitive attributes (Dwork et al., 2012). Promoting demographic parity can lead to poor performance, especially if the true outcome is not independent of the sensitive attributes. To remedy this, Hardt et al. (2016) proposed *equalized odds* to ensure that the predicted target is conditionally independent of the sensitive attributes given the true label. A further relaxed version of this notion is *equal opportunity* which is satisfied if predicted target is conditionally independent of sensitive attributes given that the true label is in an advantaged class (Hardt et al., 2016). The inherent assumption in such conditional notions is that the true labels are fair. These notions suffer from a potential amplification of the inherent discrimination that may exist in the training data. Tackling such bias is beyond the scope of this work; cf. Kilbertus et al. (2020) and Bechavod et al. (2019).

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

| Reference | NB target | NB attrib. | NB code | Fairness notion dp | Fairness notion eod | Fairness notion eop | Beyond logistic | Stoch. alg. (unbiased**) | Converg. (stoch.) |
|---|---|---|---|---|---|---|---|---|---|
| **FERMI (this work)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ (✓) | ✓ (✓) |
| (Cho et al., 2020b) | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ (✗) | ✗ |
| (Cho et al., 2020a) | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ (✓) | ✗ |
| (Baharlouei et al., 2020) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ (✗) |
| (Rezaei et al., 2020) | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| (Jiang et al., 2020)* | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| (Mary et al., 2019) | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ (✗) | ✗ |
| (Donini et al., 2018) | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| (Zhang et al., 2018) | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ (✗) | ✗ |

Table 1: Comparison of state-of-the-art in-processing methods. **NB = non-binary,** dp = demographic parity, eod = equalized odds, eop = equal opportunity. While satisfying eod guarantees satisfying eop, an eod algorithm does not necessarily achieve a favorable tradeoff between performance and fairness violation in eop; we only credit those works that provide/implement algorithms for a given fairness notion. FERMI is the only method compatible with stochastic optimization and guaranteed convergence. The only existing baselines for non-binary classification with non-binary sensitive attributes are (Mary et al., 2019; Baharlouei et al., 2020; Cho et al., 2020b) (NB code). *We refer to the in-processing method of (Jiang et al., 2020), not their post-processing method. **We use the term "unbiased" to refer to unbiased estimation in statistical sense; it is not to be confused with bias in the fairness sense, for which we use the term discrimination.

**Measuring fairness violation.** In practice, the learner only has access to finite samples and cannot verify demographic parity, equalized odds, or equal opportunity. This has led the machine learning community to define several fairness violation metrics that quantify the degree of (conditional) independence between random variables, e.g., $L_\infty$ distance (Dwork et al., 2012; Hardt et al., 2016), mutual information (Kamishima et al., 2011; Rezaei et al., 2020; Steinberg et al., 2020; Zhang et al., 2018; Cho et al., 2020a), Pearson correlation (Zafar et al., 2017), false positive/negative rates (Bechavod & Ligett, 2017), Hilbert Schmidt independence criterion (HSIC) (Pérez-Suay et al., 2017), Rényi correlation (Mary et al., 2019; Baharlouei et al., 2020; Grari et al., 2019, 2020), and exponential Rényi mutual information (ERMI) (Mary et al., 2019). In this paper, we focus on three variants of ERMI specialized to demographic parity, equalized odds, and equal opportunity. We prove that ERMI provides an upper bound on the rest of the above existing notions of fairness violation. Consequently, a model trained to reduce ERMI will also provide guarantees on these other fairness violations. We also develop a stochastic estimator for ERMI that is compatible with large-scale stochastic optimization, and use it as a regularizer in within ERM, and call it FERMI. We theoretically show that FERMI is convergent, and empirically demonstrate that it outperforms all other state-of-the-art baselines, including (Mary et al., 2019) which solves the same objective as FERMI.

**Related work & contributions.** Fairness-promoting machine learning algorithms can be categorized in three main classes: *pre-processing*, *post-processing*, and *in-processing* methods. Pre-processing algorithms (Feldman et al., 2015; Zemel et al., 2013; Calmon et al., 2017b) transform the biased data features to a new space in which the labels and sensitive attributes are statistically independent. This transform is oblivious to the training procedure. Post-processing approaches (Hardt et al., 2016; Pleiss et al., 2017) mitigate the discrimination of the classifier by altering the the final decision. In-processing approaches focus on the training procedure and impose the notions of fairness as constraints or regularization terms in the training procedure. Several regularization-based methods are proposed in the literature to promote fairness in decision-trees (Kamiran et al., 2010; Raff et al., 2018; Aghaei et al., 2019), support vector machines (Donini et al., 2018), neural networks (Grari et al., 2020; Cho et al., 2020b), or (logistic) regression models (Zafar et al., 2017; Berk et al., 2017; Taskesen et al., 2020; Chzhen & Schreuder, 2020; Baharlouei et al., 2020; Jiang et al., 2020; Grari et al., 2019). While in-processing approaches generally give rise to better tradeoffs between fairness violation and performance, existing approaches are mostly incompatible with large-scale stochastic optimization. This paper addresses this problem. See below for a summary of our contributions and Table 1 for a summary of the main differences between FERMI and existing in-processing methods.

1. We analyze a notion of fairness violation called ERMI. We show that ERMI is a stronger notion of fairness violation than all existing notions. Therefore, a model that ensures small ERMI violation is guaranteed to have small fairness violation with respect to all other notions as well.

2. We formulate an empirical objective, called FERMI objective, for using ERMI as a regularizer with empirical risk minimization. We propose a solver for FERMI, which is the first stochastic in-processing fairness algorithm with guaranteed convergence. The existing stochastic fairness

algorithms by Zhang et al. (2018); Mary et al. (2019); Cho et al. (2020a,b) are not guaranteed to converge.

3. We demonstrate through extensive numerical experiments that FERMI achieves superior fairness-accuracy tradeoff curves against all comparable baselines, even when fairness violation is measured in terms of commonly used $L_\infty$ (for demographic parity, equalized odds, and equal opportunity). In particular, the performance gap is very large when minibatch size is small (as is practically necessary for large-scale problems), and the number of sensitive attributes is large.

## 2 Fairness notions: demographic parity, equalized odds, equal opportunity

In this section, we state a notion of fairness that generalizes demographic parity, equalized odds, and equal opportunity fairness definitions (the three notions considered in this paper). This will be convenient for presenting our theoretical results. Consider a learner who trains a model to make a prediction, $\widehat{Y}$, e.g., whether or not to extend a loan, supported on $\mathcal{Y}$ which can be discrete or continuous. The prediction is made using a set of features, $\mathbf{X}$, e.g., financial history features. We assume that there is a set of discrete sensitive attributes, $S$, e.g., race and sex, supported on $\mathcal{S}$, associated with each sample. Further, let $\mathcal{A} \subseteq \mathcal{Y}$ denote an advantaged outcome class, e.g., the outcome where a loan is extended.

**Definition 1** ($(Z, \mathcal{Z})$-fairness)**.** *Given a random variable $Z$, let $\mathcal{Z}$ be a subset of values that $Z$ can take. We say that a learning machine satisfies $(Z, \mathcal{Z})$-fairness if for every $z \in \mathcal{Z}$, $\widehat{Y}$ is conditionally independent of $S$ given $Z = z$, i.e. $\forall \widehat{y} \in \mathcal{Y}, s \in \mathcal{S}, z \in \mathcal{Z}$, $p_{\widehat{Y},S|Z}(\widehat{y}, s|z) = p_{\widehat{Y}|Z}(\widehat{y}|z)p_{S|Z}(s|z)$.*

$(Z, \mathcal{Z})$-fairness includes the popular demographic parity, equalized odds, and equal opportunity notions of fairness as special cases:

1. $(Z, \mathcal{Z})$-fairness recovers demographic parity (Dwork et al., 2012) if $Z = 0$ and $\mathcal{Z} = \{0\}$. In this case, conditioning on $Z$ has no effect, and hence $(0, \{0\})$ fairness is equivalent to the independence between $\widehat{Y}$ and $S$ (see Definition 6, Appendix A).
2. $(Z, \mathcal{Z})$-fairness recovers equalized odds (Hardt et al., 2016) if $Z = Y$ and $\mathcal{Z} = \mathcal{Y}$. In this case, $Z \in \mathcal{Z}$ is trivially satisfied. Hence, conditioning on $Z$ is equivalent to conditioning on $Y$, which recovers the equalized odds notion of fairness, i.e., conditional independence of $\widehat{Y}$ and $S$ given $Y$ (see Definition 7, Appendix A).
3. $(Z, \mathcal{Z})$-fairness recovers equal opportunity (Hardt et al., 2016) if $Z = Y$ and $\mathcal{Z} = \mathcal{A}$. This is also similar to the previous case with $\mathcal{Y}$ replaced with $\mathcal{A}$ (see Definition 8, Appendix A).

Note that verifying $(Z, \mathcal{Z})$-fairness requires having access to the joint distribution of random variables $(Z, \widehat{Y}, S)$. This joint distribution is unavailable to the learner in the context of machine learning, and hence the learner would resort to empirical estimation of the amount of violation of independence, measured through some divergence. See (Williamson & Menon, 2019) for a related discussion.

## 3 Measuring fairness violation using exponential Rényi mutual information

Most existing fairness violations can be viewed as a (conditional) $f$-divergence between the joint distribution of sensitive attributes and predicted targets, $p_{\widehat{Y},S|Z}$, and the Kronecker proudct of the marginals, $p_{\widehat{Y}|Z} \otimes p_{S|Z}$. In this section, we focus on ERMI and show that several existing fairness violations are upper bounded by ERMI. For brevity, we present all definitions and results $(Z, \mathcal{Z})$.

**Definition 2** (ERMI – exponential Rényi mutual information)**.** *We define the exponential Rényi mutual information between $\widehat{Y}$ and $S$ given $Z \in \mathcal{Z}$ as*

$$D_R(\widehat{Y}; S|Z \in \mathcal{Z}) := \mathbb{E}_{Z,\widehat{Y},S}\left\{ \frac{p_{\widehat{Y},S|Z}(\widehat{Y}, S|Z)}{p_{\widehat{Y}|Z}(\widehat{Y}|Z)p_{S|Z}(S|Z)} \middle| Z \in \mathcal{Z} \right\} - 1. \tag{ERMI}$$

In Appendix B, we unravel the definition for the special cases of interest corresponding to demographic parity, equalied odds, and equal opportunity. We also discuss that ERMI is the $\chi^2$-divergence (which is an $f$-divergence) between the joint distribution, $p_{\widehat{Y},S|Z}$, and the Kronecker product of marginals, $p_{\widehat{Y}|Z} \otimes p_{S|Z}$ (Calmon et al., 2017a). In particular, ERMI is non-negative, and zero if and only if $(Z, \mathcal{Z})$-fairness is satisfied. In the context of algorithmic fairness, ERMI was first used by Mary et al. (2019) as a regularizer. We will provide a new stochastic solver/estimator for ERMI, which theoretically converges and empirically outperforms the one by Mary et al. (2019).

**Definition 3** (Rényi mutual information (Rényi, 1961)). *Let the Rényi mutual information of order $\alpha > 1$ between random variables $\widehat{Y}$ and $S$ given $Z \in \mathcal{Z}$ be defined as:*

$$I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z}) := \frac{1}{\alpha - 1} \log \left( \mathbb{E}_{Z,\widehat{Y},S} \left\{ \left( \frac{p_{\widehat{Y},S|Z}(\widehat{Y}, S|Z)}{p_{\widehat{Y}|Z}(\widehat{Y}|Z) p_{S|Z}(S|Z)} \right)^{\alpha-1} \middle| Z \in \mathcal{Z} \right\} \right), \tag{RMI}$$

*which generalizes Shannon mutual information*

$$I_1(\widehat{Y}; S|Z \in \mathcal{Z}) := \mathbb{E}_{Z,\widehat{Y},S} \left\{ \log \left( \frac{p_{\widehat{Y},S|Z}(\widehat{Y}, S|Z)}{p_{\widehat{Y}|Z}(\widehat{Y}|Z) p_{S|Z}(S|Z)} \right) \middle| Z \in \mathcal{Z} \right\}, \tag{MI}$$

*and recovers it as $\lim_{\alpha \to 1^+} I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z}) = I_1(\widehat{Y}; S|Z \in \mathcal{Z})$.*

Note that $I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z}) \geq 0$ with equality if and only if $(Z, \mathcal{Z})$-fairness is satisfied.

**Theorem 1** (ERMI is stronger than Shannon mutual information). *We have*

$$0 \leq I_1(\widehat{Y}; S|Z \in \mathcal{Z}) \leq I_2(\widehat{Y}; S|Z \in \mathcal{Z}) \leq e^{I_2(\widehat{Y}; S|Z \in \mathcal{Z})} - 1 = D_R(\widehat{Y}; S|Z \in \mathcal{Z}). \tag{1}$$

All proofs are relegated to the appendix. Theorem 1 establishes that ERMI is a stronger measure of fairness violation in the sense that driving it to zero would also bound the Shannon mutual information, which is used for promoting fairness in recent literature (Cho et al., 2020a). It also shows that ERMI is exponentially related to the Rényi mutual information of order 2.

**Definition 4** (Rényi correlation (Hirschfeld, 1935; Gebelein, 1941; Rényi, 1959)). *Let $\mathcal{F}$ and $\mathcal{G}$ be the set of measurable functions such that for random variables $\widehat{Y}$ and $S$, $\mathbb{E}_{\widehat{Y}}\{f(\widehat{Y}; z)\} = \mathbb{E}_S\{g(S; z)\} = 0$, $\mathbb{E}_{\widehat{Y}}\{f(\widehat{Y}; z)^2\} = \mathbb{E}_S\{g(S; z)^2\} = 1$, for all $z \in \mathcal{Z}$. Rényi correlation is:*

$$\rho_R(\widehat{Y}, S|Z \in \mathcal{Z}) := \sup_{f,g \in \mathcal{F} \times \mathcal{G}} \mathbb{E}_{Z,\widehat{Y},S} \left\{ f(\widehat{Y}; Z) g(S; Z) \middle| Z \in \mathcal{Z} \right\}. \tag{RC}$$

Rényi correlation generalizes Pearson correlation,

$$\rho(\widehat{Y}, S|Z \in \mathcal{Z}) := \mathbb{E}_Z \left\{ \frac{\mathbb{E}_{\widehat{Y},S}\{\widehat{Y}S|Z\}}{\sqrt{\mathbb{E}_{\widehat{Y}}\{\widehat{Y}^2|Z\} \mathbb{E}_S\{S^2|Z\}}} \middle| Z \in \mathcal{Z} \right\}, \tag{PC}$$

to capture nonlinear dependencies between the random variables by finding functions of random variables that maximize the Pearson correlation coefficient between the random variables. In fact, it is true that $\rho_R(\widehat{Y}, S|Z \in \mathcal{Z}) \geq 0$ with equality if and only if $(Z, \mathcal{Z})$-fairness is satisfied. Rényi correlation has gained popularity as a measure of fairness violation (Mary et al., 2019; Baharlouei et al., 2020; Grari et al., 2020). Rényi correlation is also upper bounded by ERMI. The following result has already been shown by Mary et al. (2019) and we present it for completeness.

**Theorem 2** (ERMI is stronger than Rényi correlation). *We have*

$$0 \leq |\rho(\widehat{Y}, S|Z \in \mathcal{Z})| \leq \rho_R(\widehat{Y}, S|Z \in \mathcal{Z}) \leq D_R(\widehat{Y}; S|Z \in \mathcal{Z}), \tag{2}$$

*and if $|\mathcal{S}| = 2$, $D_R(\widehat{Y}; S|Z \in \mathcal{Z}) = \rho_R(\widehat{Y}, S|Z \in \mathcal{Z})$.*

**Definition 5** ($L_q$ fairness violation). *We define the $L_q$ fairness violation for $q \geq 1$ by:*

$$L_q(\widehat{Y}, S|Z \in \mathcal{Z}) := \mathbb{E}_Z \left\{ \left( \int_{\widehat{y} \in \mathcal{Y}_0} \sum_{s \in \mathcal{S}_0} \left| p_{\widehat{Y},S|Z}(\widehat{y}, s|Z) - p_{\widehat{Y}|Z}(\widehat{y}|Z) p_{S|Z}(s|Z) \right|^q dy \right)^{\frac{1}{q}} \middle| Z \in \mathcal{Z} \right\}. \tag{Lq}$$

Note that $L_q(\widehat{Y}, S|Z \in \mathcal{Z}) = 0$ if and only if $(Z, \mathcal{Z})$-fairness is satisfied. In particular, $L_\infty$ fairness violation recovers demographic parity violation (Kearns et al., 2018, Definition 2.1) if we let $\mathcal{Z} = \{0\}$ and $Z = 0$. It also recovers equal opportunity violation (Hardt et al., 2016) if $\mathcal{Z} = \mathcal{A}$ and $Z = Y$.

**Theorem 3** (ERMI is stronger than $L_\infty$ fairness violation). *Let $\widehat{Y}$ be a discrete or continuous random variable, and $S$ be a discrete random variable supported on a finite set. Then for any $q \geq 1$,*

$$0 \leq L_q(\widehat{Y}, S|Z \in \mathcal{Z}) \leq \sqrt{D_R(\widehat{Y}, S|Z \in \mathcal{Z})}. \tag{3}$$

The above theorem says that if a method controls ERMI value for imposing fairness, then $L_\infty$ violation is controlled. In particular, the variant of ERMI that is specialized to demographic parity also controls $L_\infty$ demographic parity violation (Kearns et al., 2018). The variant of ERMI that is specialized to equal opportunity also controls the $L_\infty$ equal opportunity violation (Hardt et al., 2016). While our algorithm uses ERMI as a regularizer, in our experiments, we measure fairness violation through the more commonly used $L_\infty$ violation. Despite this, we show that our approach leads to better tradeoff curves between fairness violation and performance.

**Remark.** The bounds in Theorems 1-3 are not tight in general, but this is not of practical concern. They show that bounding ERMI is sufficient because any model that achieves small ERMI is guaranteed to satisfy any other fairness violation. This makes ERMI an effective regularizer for promoting fairness. In fact, in Sec. 5, we see that the proposed algorithm, FERMI, achieves the best tradeoffs between fairness violation and performance across state-of-the-art baselines.

# 4 FERMI: fair empirical risk minimization through ERMI regularization

Our goal is to train a model that balances fairness and accuracy objectives. To this end, we introduce fair risk minimization through exponential Rényi mutual information framework defined below:[1]

$$\min_{\boldsymbol{\theta}} \Big\{ \text{FRMI}(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{X},Y,S} \left\{ \ell\big(\mathbf{X}, Y; \boldsymbol{\theta}\big) \right\} + \lambda D_R\big(\widehat{Y}(\mathbf{X};\boldsymbol{\theta}); S\big) \Big\}, \qquad \text{(FRMI obj.)}$$

where $\ell$ denotes the loss function, such as $L_2$ loss or cross entropy loss; $\lambda > 0$ is a scalar balancing the accuracy versus fairness objectives; $D_R\big(\widehat{Y}(\mathbf{X};\boldsymbol{\theta}); S\big)$ is the notion of ERMI given in Eq. (ERMI) particularized to demographic parity (see Eq. (5)); and $\widehat{Y}(\mathbf{X};\boldsymbol{\theta})$ is the output of the learned model (e.g., the output of a classification or a regression task, or the cluster number in a clustering task). While $\widehat{Y}(\mathbf{X};\boldsymbol{\theta})$ inherently depends on $\mathbf{X}$ and $\boldsymbol{\theta}$, in the rest of this paper, we sometimes leave the dependence of $\widehat{Y}$ on $\mathbf{X}$ and/or $\boldsymbol{\theta}$ implicit for brevity of notation. Notice that we have also left the dependence of the loss on the predicted outcome $\widehat{Y}$ implicit.

In practice, the true joint distribution of $(\mathbf{X}, S, Y, \widehat{Y})$ is unknown and we only have $N$ samples at our disposal, making it impossible to solve FRMI. Let $\{\mathbf{x}_i, s_i, y_i, \widehat{y}_i(\mathbf{x}_i;\boldsymbol{\theta})\}_{i\in[N]}$ denote the features, sensitive attributes, targets, and the predictions of the model parameterized by $\boldsymbol{\theta}$ for these samples. Mary et al. (2019) considered the same objective Eq. (FRMI obj.), and tried to empirically solve it through a kernel approximation. We propose a completely different approach to solving this problem: fair empirical risk minimization via exponential Rényi mutual information (FERMI). FERMI results in a provably convergent algorithm, and empirically outperforms the algorithm by Mary et al. (2019). It is straightforward to derive an unbiased estimate for $\mathbb{E}_{\mathbf{X},Y,S} \left\{ \ell\big(\mathbf{X}, Y; \boldsymbol{\theta}\big) \right\}$ through the empirical risk, e.g., $\frac{1}{|B|}\sum_{i\in B} \ell\big(\mathbf{x}_i, y_i; \boldsymbol{\theta}\big)$ where $B \subseteq [N]$ is a random minibatch of data points. However, estimating $D_R(\widehat{Y}, S)$ in the objective function in Eq. (FRMI obj.) is more difficult. In what follows, we present our approach to deriving an *unbiased stochastic estimator* of $D_R(\widehat{Y}, S)$ given a random batch of data points $B$. The following theorem is the key tool we use to obtain an unbiased estimator:

**Theorem 4.** *For discrete random variables $\widehat{Y} = \widehat{Y}(\mathbf{X};\boldsymbol{\theta})$ and $S$ where $\widehat{Y} \in [m], S \in [k]$, we have*

$$D_R(\widehat{Y}; S) = \max_{W \in \mathbb{R}^{k\times m}} \Big\{ -\operatorname{Tr}(W P_{\widehat{y}} W^T) + 2\operatorname{Tr}(W P_{\widehat{y},s} P_s^{-1/2}) - 1 \Big\}, \tag{4}$$

*where* $P_{\widehat{y}} = \operatorname{diag}(p_{\widehat{Y}}(1), \ldots, p_{\widehat{Y}}(m))$, $P_s = \operatorname{diag}(p_S(1), \ldots, p_S(k))$, *and*

$$P_{\widehat{y},s} = \begin{pmatrix} p_{\widehat{Y},S}(1,1) & \ldots & p_{\widehat{Y},S}(1,k) \\ \vdots & \ddots & \vdots \\ p_{\widehat{Y},S}(m,1) & \ldots & p_{\widehat{Y},S}(m,k) \end{pmatrix}.$$

Let $\widehat{\mathbf{Y}},\ \widehat{\mathbf{y}}_i \in \{0,1\}^m$ and $\mathbf{S},\ \mathbf{s}_i \in \{0,1\}^k$ be the one-hot encodings of $\widehat{Y}, \widehat{y}_i$ and $S, s_i$, respectively. Then, the above theorem implies that we can compute an unbiased estimate of Eq. (FRMI obj.):

[1] In this section, we present all results in the context of $Z = 0$ and $\mathcal{Z} = \{0\}$ (demographic parity), leaving off all conditional expectations for clarity of presentation. The results are readily generalized for general $(Z, \mathcal{Z})$ by using $D_R(\widehat{Y}, S|Z \in \mathcal{Z})$ in Eq. (FRMI obj.)); we have used the resulting algorithms for empirical experiments.

**Lemma 1** (Unbiased estimator of ERMI). *Let* $(\mathbf{X}, S, Y, \widehat{Y}(\mathbf{X}; \boldsymbol{\theta}))$ *be a random draw from* $P_{\mathbf{X},S,Y,\widehat{Y}}$. *Further, let*

$$\psi(\mathbf{X}, S, Y, \widehat{Y}; \boldsymbol{\theta}, W) := -\operatorname{Tr}(W\widehat{\mathbf{Y}}(\mathbf{X};\boldsymbol{\theta})\widehat{\mathbf{Y}}^T(\mathbf{X};\boldsymbol{\theta})W^T) + 2\operatorname{Tr}(W\widehat{\mathbf{Y}}(\mathbf{X};\boldsymbol{\theta})\mathbf{S}^T P_s^{-1/2}) - 1.$$

*Then,* $\max_{W\in\mathbb{R}^{k\times m}} \psi(\mathbf{X}, S, Y, \widehat{Y}; \boldsymbol{\theta}, W)$ *is an unbiased estimator of ERMI in Eq.* (FRMI obj.)*, i.e.,*

$$\mathbb{E}_{\mathbf{X},S,Y}\left\{\max_{W\in\mathbb{R}^{k\times m}} \psi(\mathbf{X}, S, Y, \widehat{Y}; \boldsymbol{\theta}, W)\right\} = D_R(\widehat{Y}(\mathbf{X};\boldsymbol{\theta}); S).$$

The stochastic estimator, $\psi(\mathbf{X}, S, Y, \widehat{Y}; \boldsymbol{\theta}, W)$, in Lemma 1 requires the knowledge of $P_s$, and computation of $P_s^{-1/2}$. This can be estimated with high fidelity (for small to moderate sensitive set) through a single initial pass over the entire dataset in practice. Hence, we consider it to be known. Now, we are equipped to state the empirical objective function that we solve in this paper:

$$\min_{\boldsymbol{\theta}} \max_{W\in\mathbb{R}^{k\times m}} \left\{\text{FERMI}(\boldsymbol{\theta}, W) := \frac{1}{N}\sum_{i\in[N]} [\ell(\mathbf{x}_i, y_i; \boldsymbol{\theta}) + \lambda\psi_i(\boldsymbol{\theta}, W)]\right\}, \qquad \text{(FERMI obj.)}$$

where

$$\psi_i(\boldsymbol{\theta}, W) := -\operatorname{Tr}(W\widehat{\mathbf{y}}_i(\mathbf{x}_i;\boldsymbol{\theta})\widehat{\mathbf{y}}_i^T(\mathbf{x}_i;\boldsymbol{\theta})W^T) + 2\operatorname{Tr}(W\widehat{\mathbf{y}}_i(\mathbf{x}_i;\boldsymbol{\theta})\mathbf{s}_i^T P_s^{-1/2}) - 1.$$

In particular, Lemma 1 says that, for any $N$, Eq. (FERMI obj.) (and its gradients) is an *unbiased* and *consistent* estimator of the Eq. (FRMI obj.) objective function (and its gradients) by an empirical average over the minibatch. This is in contrast to the density estimation methods used by Mary et al. (2019) and Baharlouei et al. (2020), which are biased but consistent. We will see in the experiments that the unbiased estimator empirically offers large performance improvements.

This observations leads us to deriving a stochastic algorithm, presented in Algorithm 1, which is guaranteed to converge for any batch size $1 \le |B| \le N$ since the stochastic gradients are unbiased.

**Algorithm 1** (FERMI Algorithm). Two-Time Scale SGDA for solving FERMI objective

1: **Input**: $\boldsymbol{\theta}^0 \in \mathbb{R}^{d_\theta}$, $W^0 \in \mathcal{W} \subset \mathbb{R}^{k\times m}$, step-sizes $(\eta_\theta, \eta_w)$, mini-batch $B \subseteq [N]$, fairness parameter $\lambda \ge 0$, iteration number $R$.
2: **for** $t = 0, 1, \ldots, R$ **do**
3: Draw a mini-batch $B$ of data points $\{(\mathbf{x}_i, s_i, y_i)\}_{i\in B}$
4: Set $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \frac{\eta_\theta}{|B|}\sum_{i\in B}[\nabla_\theta \ell(\mathbf{x}_i, y_i; \boldsymbol{\theta}^t) + \lambda\nabla_\theta\psi_i(\boldsymbol{\theta}^t, W^t)]$.
5: Set $W^{t+1} \leftarrow \Pi_{\mathcal{W}}\Big(W^t + \frac{2\lambda\eta_w}{|B|}\sum_{i\in B}\Big[-W\widehat{\mathbf{y}}_i(\mathbf{x}_i;\boldsymbol{\theta}^t)\widehat{\mathbf{y}}_i^T(\mathbf{x}_i;\boldsymbol{\theta}^t) + P_s^{-1/2}\mathbf{s}_i\widehat{\mathbf{y}}_i^T(\mathbf{x}_i;\boldsymbol{\theta}^t)\Big]\Big)$
6: **end for**
7: Pick $\hat{t}$ uniformly at random from $\{1, \ldots, R\}$.
8: **Return:** $\boldsymbol{\theta}^{\hat{t}}$.

**Theorem 5.** *(Informal statement) Algorithm 1 converges to the set of $\epsilon$-first order stationary points of the Eq.* (FERMI obj.) *objective in* $O(\frac{1}{\epsilon^4})$ *iterations (stochastic gradient evaluations).*

The formal statement of this theorem can be found in Theorem 10 in Appendix D. A faster convergence rate of $O(\frac{1}{\epsilon^3})$ could be obtained by using the (more complicated) SREDA method of Luo et al. (2020) instead of SGDA to solve FERMI objective. We omit the details here. In the next section, we numerically evaluate the performance FERMI algorithm in several numerical experiments.

# 5 Numerical experiments

## 5.1 Binary classification and binary sensitive attribute

For our first set of experiments, we evaluate the fairness-accuracy tradeoffs of FERMI in binary classification problems with a binary sensitive attribute. This is a common setup, so we are able to compare against many existing baseline methods (Zafar et al., 2017; Feldman et al., 2015; Kamishima et al., 2011; Jiang et al., 2020; Hardt et al., 2016; Baharlouei et al., 2020; Rezaei et al., 2020; Donini et al., 2018; Cho et al., 2020b). We run experiments on three data sets: Adult, German Credit, and COMPAS. To implement FERMI, we train a logistic regression model (same model for all baselines) with an ERMI regularizer. Details about the datasets and experiments can be found in Appendix E.

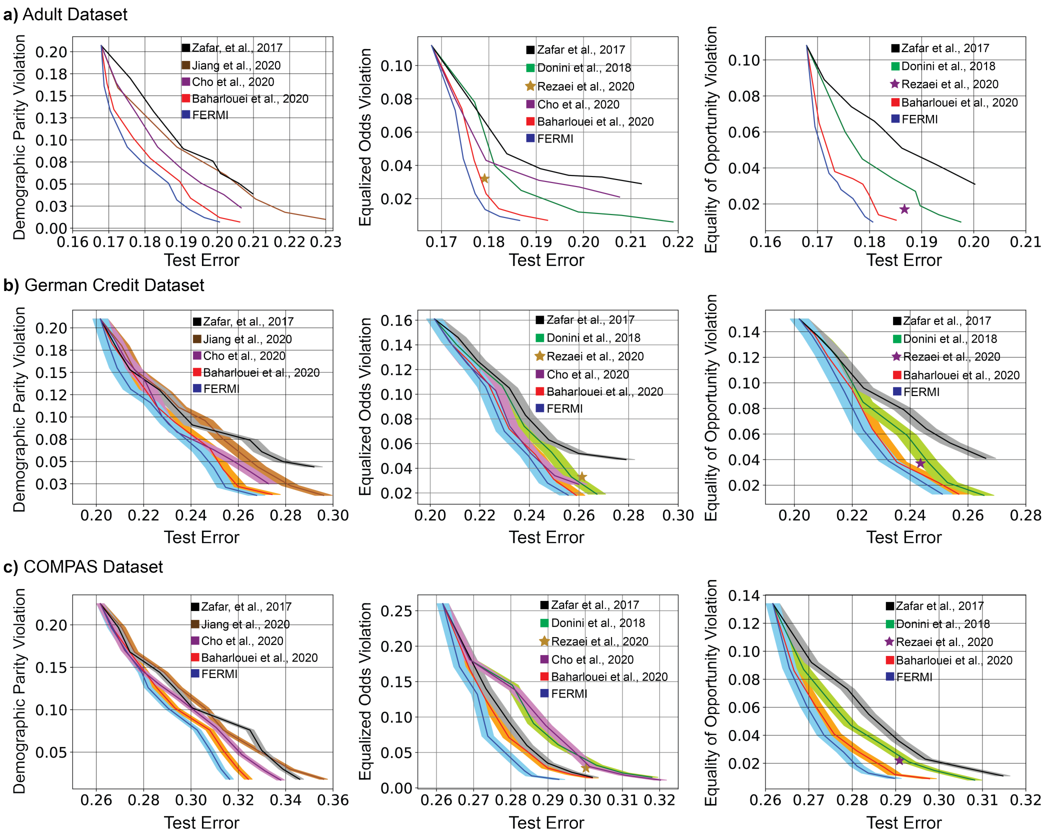


Figure 1: Binary classification with binary sensitive attribute using logistic regression. Tradeoff of fairness violation vs. test error for state-of-the-art fair classifiers on German Credit, Adult, and COMPAS datasets. FERMI offers the best fairness vs. accuracy tradeoff curve in all experiments against all baselines. Rezaei et al. (2020) only allow for a single output and do not yield a tradeoff curve. Further, the algorithms by Mary et al. (2019) and Baharlouei et al. (2020) are equivalent in this binary setting and shown by the red curve. FERMI, Mary et al. (2019) and Baharlouei et al. (2020) try to empirically solve the same risk function Eq. (FRMI obj.). However, the empirical formulation used by FERMI, Eq. (FERMI obj.) and its solver result in a better performance even-though we are using a full-batch for all baselines in this experiment.

In Fig. 1, we report the fairness violation vs. test error, for three notions of fairness: demographic parity, equalized odds, and equal opportunity. We have only included in-processing methods, which outperform pre-processing and post-processing methods. Complete experimental results are included in the appendix. We measure fairness violation through conditional demographic parity $L_\infty$ violation (Definition 9), conditional equal opportunity $L_\infty$ violation (Definition 10) and its generalization, conditional equalized odds violation. As can be seen, FERMI offers a fairness-accuracy tradeoff curve that dominates all existing state-of-the-art baselines in each experiment and with respect to each notion of fairness. This demonstrates the efficacy of having a strong regularizer such as ERMI: by enforcing small ERMI violation, our model simultaneously achieves small fairness violation with respect to these other notions which are upper bounded by ERMI.

It is noteworthy that the empirical objective function of Mary et al. (2019) and Baharlouei et al. (2020) is exactly the same in this setting, and their algorithms also coincide to the red curve in Fig. 1.[2] Additionally, like FERMI, they are trying to empirically solve Eq. (FRMI obj.), albeit using different estimation techniques, i.e., their empirical objective is different from Eq. (FERMI obj.). This demonstrates the effectiveness of our empirical formulation (FERMI obj.) – which is both unbiased and consistent whereas theirs is biased. It also shows the effectiveness of our solver (Algorithm 1) even-though we are using all baselines in full batch mode in this experiment. In the following experiments, we will demonstrate that using smaller batch sizes results in much more pronounced advantages of FERMI over these baselines.

[2]Exponential Rényi mutual information is equal to Rényi correlation for binary targets and/or binary sensitive attributes (see Theorem 2), which is the setting of all experiments in Sec. 5.1.

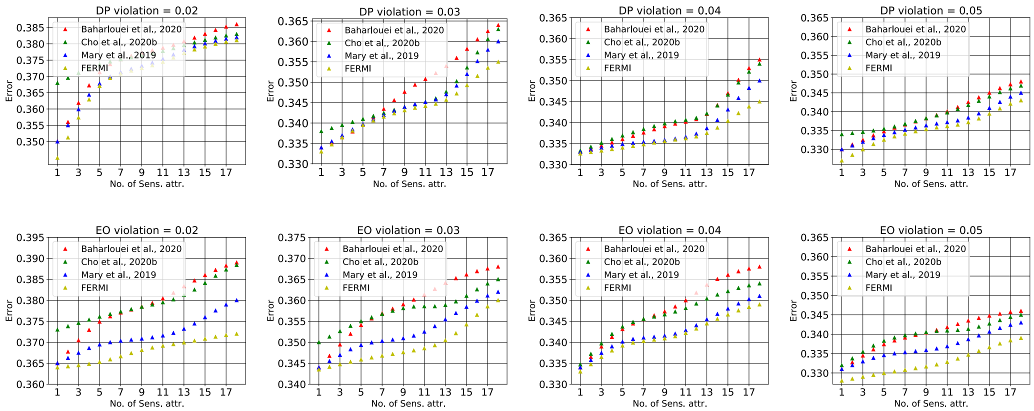


Figure 2: Comparison between FERMI, Mary et al. (2019), Baharlouei et al. (2020), and Cho et al. (2020b) on Communities dataset. (Mary et al., 2019) outperforms (Baharlouei et al., 2020; Cho et al., 2020b), which we believe could be attributed to the effectiveness of ERMI as a regularizer. FERMI outperforms Mary et al. (2019), which we attribute to our empirical formulation of ERMI and the effectiveness of its solver, given that we try to empirically solve the same risk function with different formulations.

### 5.2 Non-binary fair classification with a non-binary sensitive attribute

Next, we consider a non-binary classification problem with non-binary sensitive set. In this case, we consider the Communities and Crime dataset, which has 18 binary sensitive attributes in total, and we pick a subset of $1, 2, 3, \ldots, 18$ sensitive attributes out of those for our experiments, which corresponds to $|\mathcal{S}| \in \{2, 4, 8, \ldots, 2^{18}\}$. We discretize the target into three classes $\{\text{high}, \text{medium}, \text{low}\}$. The only baselines that we are aware of that can handle non-binary classification with non-binary sensitive attributes are (Mary et al., 2019), (Baharlouei et al., 2020), (Cho et al., 2020b), (Cho et al., 2020a), and (Zhang et al., 2018). We used the publicly available implementations of (Baharlouei et al., 2020) and (Cho et al., 2020b) and extended their binary classification algorithms to the non-binary setting.

The results are presented in Fig. 2, where we use conditional demographic parity $L_\infty$ violation (Definition 9) and conditional equal opportunity $L_\infty$ violation (Definition 10) as the fairness violation notions for the two experiments. For all baselines, test error increases as the number of sensitive attributes increases. As can be seen, compared to the baselines, FERMI offers the most favorable test error vs. fairness violation tradeoffs, particularly as the number of sensitive attributes increases and for the more stringent fairness violation levels, e.g., $0.02$.

### 5.3 Domain generalization through FERMI

In our last experiment, our goal is to showcase the efficacy of FERMI in stochastic optimization with neural network approximation. For this experiment, we consider the Color MNIST dataset (Li & Vasconcelos, 2019), where all 60,000 training MNIST digits are colored with different colors drawn from a class conditional Gaussian distribution with variance $\sigma$ around a certain average color for each digit, while the test set remains black and white. Li & Vasconcelos (2019) show that as $\sigma \to 0$, a convolutional network model overfits significantly to each digit's color on the training set, and achieves vanishing training accuracy. However, the learned representation does not generalize to the regular black and white test set, in absence of the spurious correlation between digits and color.

Conceptually, the goal of the classifier in this problem is to achieve high classification accuracy with predictions that are independent of the color of the digit. We view color as the sensitive attribute in this experiment, and apply fairness baselines for the demographic parity notion of fairness. One would expect that by promoting such independence through a fairness regularizer generalization would improve (i.e. lower test error on the black and white test set), at the cost of increased training error (on the colored training set). We compare against Mary et al. (2019), Baharlouei et al. (2020), and Cho et al. (2020b) as baselines in this experiment.

The results of this experiment are as illustrated in Fig. 3. The details about the dataset and experimental setup is provided in Appendix E. In the left panel, we see that with no regularization ($\lambda = 0$); the test error is around 80%. As $\lambda$ increases, all methods achieve smaller test error while training error increases. We also observe that FERMI offers the best test error in this setup. In the right panel, we observe that decreasing the batch size results in significantly worse generalization for all three

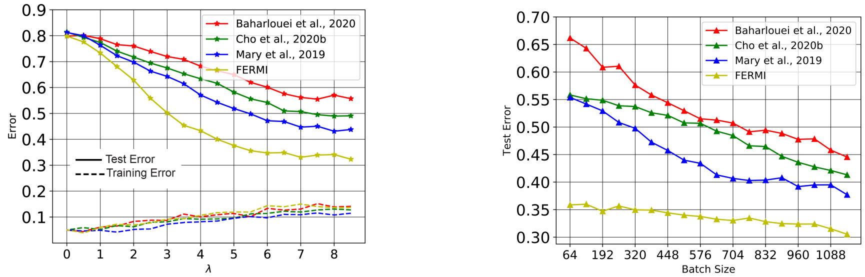


Figure 3: Domain generalization on Color MNIST (Li & Vasconcelos, 2019) using in-process fair algorithms for demographic parity. **Left panel:** The dashed line is the training error and the solid line is test error. As $\lambda$ increases, fairness regularization results in a learned representation that is less dependent on color; hence training error increases while test error decreases (all algorithms reach a plateau around $\lambda = 8$). We use $|B| = 512$ for all baselines. **Right panel:** We plot test error vs. batch size using an optimized value of $\lambda$ for each algorithm selected via a validation set. The performance of all baselines drops 10-20% as batch size becomes small whereas FERMI is relatively insensitive to batch size.

baselines considered (due to their biased estimators for the regularizer). However, the impact is much less on FERMI. In particular, the performance gap between FERMI and other baselines is more than 20% for $|B| = 64$. Finally, FERMI with minibatch size $|B| = 64$ still outperforms all other baselines with $|B| > 1,000$. Finally, notice that the test error achieved by FERMI when $\sigma = 0$ is $\sim 30\%$, as compared to more than $50\%$ obtained using REPAIR (Li & Vasconcelos, 2019) for $\sigma \leq 0.05$.

# 6 Discussion & concluding remarks

In this paper, we studied three variants of a notion of fairness violation, called exponential Rényi mutual information (ERMI), developed for demographic parity, equalized odds, and equal opportunity notions of fairness. We showed that ERMI is a strong fairness violation divergence providing upper bound guarantees on other popular violation divergences, namely Shannon mutual information, Rényi mutual information (Theorem 1), Pearson correlation, Rényi correlation (Theorem 2) , and $L_q$ distance violation (Theorem 3).

We derived an unbiased estimator for ERMI (Lemma 1), based on which we formulated an empirical objective (FERMI obj.) for solving fair empirical risk minimization with ERMI regularization to balance performance and fairness. We provided a stochastic algorithm for solving FERMI (Algorithm 1) and proved its convergence (Theorem 5); for non-binary sensitive attributes, non-binary target variables, regardless of the batch size. From an experimental perspective, we showed that FERMI leads to better fairness-accuracy tradeoffs than all of the state-of-the-art baselines on a wide variety of binary and non-binary classification tasks (for demographic parity, equalized odds, and equal opportunity). We also showed that these benefits are particularly significant when the number of sensitive attributes grows or the batch size is small. In particular, we observed that FERMI consistently outperforms Mary et al. (2019) (which tries to empirically solve the same objective Eq. (FRMI obj.)) by up to 20% when the batch size is small, suggesting that the unbiasedness of the FERMI estimator is essential in achieving good empirical performance.

There are several possible explanations for the superior empirical performance of FERMI compared to baselines. One possible reason is that the objective function Eq. (FERMI obj.) is easier to optimize than the objectives of competing in-processing methods: ERMI is smooth; and in the discrete case, is equal to the trace of a matrix (see Theorem 7; appendix), which is easy to compute. Contrast this with the larger computational overhead of Rényi correlation used by Baharlouei et al. (2020), for example, which requires finding the second singular value of a matrix. Furthermore, the sample complexity of estimating Rényi mutual information of order 2 (and consequently that of ERMI) scales as $\Theta(\sqrt{|\mathcal{S}|})$ as compared to Shannon mutual information which scales as $\Theta(|\mathcal{S}|/\log|\mathcal{S}|)$ (Acharya et al., 2014). Moreover, the fact that ERMI is a stronger fairness violation seems to imply that FERMI would generalize well to other fairness notions, a hypothesis that is supported by our experimental results. Together, these facts suggest that ERMI serves as an efficient and easily optimizable proxy for these other fairness notions, making Eq. (FERMI obj.) a good surrogate objective to optimize for all three notions of fairness considered (demographic parity, equalized odds, and equal opportunity). We leave it as future work to rigorously understand which of these (or other) factors are most responsible for the favorable performance tradeoffs observed from FERMI.

# References


Acharya, J., Orlitsky, A., Suresh, A. T., and Tyagi, H. The complexity of estimating Rényi entropy. *arXiv:1408.1000v1*, 2014.

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1418–1426, 2019.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 2016.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rényi fair inference. In *ICLR*, 2020.

Bechavod, Y. and Ligett, K. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Bechavod, Y., Ligett, K., Roth, A., Waggoner, B., and Wu, Z. S. Equal opportunity in online classification with partial feedback. *arXiv preprint arXiv:1902.02242*, 2019.

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Calmon, F., Makhdoumi, A., Médard, M., Varia, M., Christiansen, M., and Duffy, K. R. Principal inertia components and applications. *IEEE Transactions on Information Theory*, 63(8):5011–5038, 2017a.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017b.

Cho, J., Hwang, G., and Suh, C. A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2521–2526. IEEE, 2020a.

Cho, J., Hwang, G., and Suh, C. A fair classifier using kernel density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Chzhen, E. and Schreuder, N. A minimax framework for quantifying risk-fairness trade-off in regression. *arXiv preprint arXiv:2007.14265*, 2020.

Cover, T. M. and Thomas, J. A. Information theory and statistics. *Elements of Information Theory*, 1 (1):279–335, 1991.

Csiszár, I. and Shields, P. C. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications*. Springer Science & Business Media, 2009.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

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

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

Fish, B., Kun, J., and Lelkes, Á. D. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Gebelein, H. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Grari, V., Ruf, B., Lamprier, S., and Detyniecki, M. Fairness-aware neural Réyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

Grari, V., Hajouji, O. E., Lamprier, S., and Detyniecki, M. Learning unbiased representations via Rényi minimization. *arXiv preprint arXiv:2009.03183*, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hirschfeld, H. O. A connection between correlation and contingency. In *Proceedings of the Cambridge Philosophical Society*, volume 31, pp. 520–524, 1935.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR, 2020.

Kamiran, F., Calders, T., and Pechenizkiy, M. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pp. 869–874. IEEE, 2010.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572, 2018.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. Fair decisions despite imperfect predictions. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 277–287. PMLR, 26–28 Aug 2020.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, Y. and Vasconcelos, N. REPAIR: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9572–9581, 2019.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv: 1906.00331v6*, 2020.

Luo, L., Ye, H., and Zhang, T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *arXiv: 2001.03724*, 2020.

Mary, J., Calauzenes, C., and El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391. PMLR, 2019.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Raff, E., Sylvester, J., and Mills, S. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 243–250, 2018.

Rényi, A. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10 (3-4):441–451, 1959.

Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

Rezaei, A., Fathony, R., Memarrast, O., and Ziebart, B. D. Fairness for robust log loss classification. In *AAAI*, pp. 5511–5518, 2020.

Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., and Caetano, T. Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*, 2020.

Sweeney, L. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*, 2013.

Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Williamson, R. and Menon, A. Fairness risk measures. In *International Conference on Machine Learning*, pp. 6786–6797. PMLR, 2019.

Witsenhausen, H. S. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

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

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendix

We provide a simple table of contents below for easier navigation of the appendix.

**CONTENTS**

## A Existing notions of fairness

Let $(Y, \widehat{Y}, \mathcal{A}, S)$ denote the true target, predicted target, the advantaged outcome class, and the sensitive attribute, respectively. We review three major notions of fairness.

**Definition 6** (demographic parity (Dwork et al., 2012))**.** *We say that a learning machine satisfies demographic parity if $\widehat{Y}$ is independent of $S$.*

**Definition 7** (equalized odds (Hardt et al., 2016))**.** *We say that a learning machine satisfies equalized odds, if $\widehat{Y}$ is conditionally independent of $S$ given $Y$.*

**Definition 8** (equal opportunity (Hardt et al., 2016))**.** *We say that a learning machine satisfies equal opportunity with respect to $\mathcal{A}$, if $\widehat{Y}$ is conditionally independent of $S$ given $Y = y$ for all $y \in \mathcal{A}$.*

Notice that the equal opportunity as defined here generalizes the definition in (Hardt et al., 2016). It recovers equalized odds if $\mathcal{A} = \mathcal{Y}$, and it recovers equal opportunity of (Hardt et al., 2016) for $\mathcal{A} = \{1\}$ in binary classification.

## B Properties and special cases of ERMI

Notice that ERMI is in fact the $\chi^2$-divergence between the conditional joint distribution, $p_{\widehat{Y},S}$, and the Kronecker product of conditional marginals, $p_{\widehat{Y}} \otimes p_S$, where the conditioning is on $Z \in \mathcal{Z}$. Further, $\chi^2$-divergence is an $f$-divergence with $f(t) = (t-1)^2$. See (Csiszár & Shields, 2004, Section 4) for a discussion. As an immediate result of this observation and well-known properties of $f$-divergences, we can state the following property of ERMI:

**Remark 6.** *$D_R(\widehat{Y}; S|Z \in \mathcal{Z}) \geq 0$ with equality if and only if for all $z \in \mathcal{Z}$, $\widehat{Y}$ and $S$ are conditionally independent given $Z = z$.*

To further clarify the definition of ERMI, especially as it relates to demographic parity, equalized odds, and equal opportunity, we will unravel the definition explicitly in a few special cases.

First, let $Z = 0$ and $\mathcal{Z} = \{0\}$. In this case, $Z \in \mathcal{Z}$ trivially holds, and conditioning on $Z$ has no effect, resulting in:

$$\begin{aligned} D_R(\widehat{Y}; S) &:= D_R(\widehat{Y}; S|Z \in \mathcal{Z})\Big|_{Z=0, \mathcal{Z}=\{0\}} \\ &= \mathbb{E}_{\widehat{Y},S}\left\{\frac{p_{\widehat{Y},S}(\widehat{Y}, S)}{p_{\widehat{Y}}(\widehat{Y})p_S(S)}\right\} - 1 \\ &= \sum_{s\in\mathcal{S}} \int_{\widehat{y}\in\mathcal{Y}} \frac{p_{\widehat{Y},S}(\widehat{y}, s) - p_{\widehat{Y}}(\widehat{y})p_S(s)}{p_{\widehat{Y}}(\widehat{y})p_S(s)} p_{\widehat{Y},S}(\widehat{y}, s) d\widehat{y}. \end{aligned} \tag{5}$$

$D_R(\widehat{Y}; S)$ is the notion of ERMI that should be used when the desired notion of fairness is demographic parity. In particular, $D_R(\widehat{Y}; S) = 0$ implies that $\chi^2$ divergence between $p_{\widehat{Y},S}$, and the Kronecker product of marginals, $p_{\widehat{Y}} \otimes p_S$ is zero. This in turn implies that $\widehat{Y}$ and $S$ are independent, which is the definition of demographic parity. We note that when $\widehat{Y}$ and $S$ are discrete, this special case ($Z = 0$ and $\mathcal{Z} = \{0\}$) of ERMI is referred to as $\chi^2$-information by Calmon et al. (2017a).

Next, we consider $Z = Y$ and $\mathcal{Z} = \mathcal{Y}$. In this case, $Z \in \mathcal{Z}$ is trivially satisfied, and hence,

$$\begin{aligned} D_R(\widehat{Y}; S|Y) &:= D_R(\widehat{Y}; S|Z \in \mathcal{Z})\Big|_{Z=Y, \mathcal{Z}=\mathcal{Y}} \\ &= \mathbb{E}_{Y,\widehat{Y},S}\left\{\frac{p_{\widehat{Y},S|Y}(\widehat{Y}, S|Y)}{p_{\widehat{Y}|Y}(\widehat{Y}|Y)p_{S|Y}(S|Y)}\right\} - 1 \\ &= \sum_{s\in\mathcal{S}} \int_{y\in\mathcal{Y}} \int_{\widehat{y}\in\mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y) - p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)}{p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)} p_{Y,\widehat{Y},S}(y, \widehat{y}, s) d\widehat{y}dy \\ &= \sum_{s\in\mathcal{S}} \int_{y\in\mathcal{Y}} \int_{\widehat{y}\in\mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y)^2}{p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)} p_Y(y) d\widehat{y}dy - 1. \end{aligned} \tag{6}$$

$D_R(\widehat{Y}; S|Y)$ should be used when the desired notion of fairness is equalized odds. In particular, $D_R(\widehat{Y}; S|Y) = 0$ directly implies the conditional independence of $\widehat{Y}$ and $S$ given $Y$.

Finally, we consider $Z = Y$ and $\mathcal{Z} = \mathcal{A}$. In this case, we have

$$\begin{aligned} D_R^{\mathcal{A}}(\widehat{Y}; S|Y) &:= D_R(\widehat{Y}; S|Z \in \mathcal{Z})\Big|_{Z=Y, \mathcal{Z}=\mathcal{A}} \\ &= \mathbb{E}_{Y,\widehat{Y},S}\left\{\frac{p_{\widehat{Y},S|Y}(\widehat{Y}, S|Y)}{p_{\widehat{Y}|Y}(\widehat{Y}|Y)p_{S|Y}(S|Y)}\middle| Y \in \mathcal{A}\right\} - 1 \\ &= \sum_{s\in\mathcal{S}} \int_{y\in\mathcal{A}} \int_{\widehat{y}\in\mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y) - p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)}{p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)} p_Y^{\mathcal{A}}(y) d\widehat{y}dy \\ &= \sum_{s\in\mathcal{S}} \int_{y\in\mathcal{A}} \int_{\widehat{y}\in\mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y)^2}{p_{\widehat{Y}|Y}(\widehat{y}|y)p_{S|Y}(s|y)} p_{\widehat{Y},S|Y}(\widehat{y}, s|y) p_Y^{\mathcal{A}}(y) d\widehat{y}dy - 1, \end{aligned} \tag{7}$$

where

$$p_Y^{\mathcal{A}}(y) := \frac{p_Y(y)}{\int_{y' \in \mathcal{A}} p_Y(y')dy'}. \tag{8}$$

This notion is what should be used when the desired notion of fairness is equal opportunity. This can be further simplified when the advantaged class is a singleton (which is the case in binary classification). If $Z = Y$ and $\mathcal{Z} = \{y\}$, then

$$\begin{aligned} D_R(\widehat{Y}; S|Y = y) &:= D_R^{\{y\}}(\widehat{Y}; S|Y) \\ &= \sum_{s \in \mathcal{S}} \int_{\widehat{y} \in \mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y) - p_{\widehat{Y}|Y}(\widehat{y}|y) p_{S|Y}(s|y)}{p_{\widehat{Y}|Y}(\widehat{y}|y) p_{S|Y}(s|y)} p_{\widehat{Y},S|Y}(\widehat{y}, s|y) d\widehat{y} \\ &= \sum_{s \in \mathcal{S}} \int_{\widehat{y} \in \mathcal{Y}} \frac{p_{\widehat{Y},S|Y}(\widehat{y}, s|y)^2}{p_{\widehat{Y}|Y}(\widehat{y}|y) p_{S|Y}(s|y)} d\widehat{y} - 1. \end{aligned} \tag{9}$$

Finally, we note that we use the notation $D_R(\widehat{Y}; S|Y)$ and $D_R(\widehat{Y}; S|Y = y)$ to be consistent with the definition of conditional mutual information in (Cover & Thomas, 1991).

## C Relations between ERMI and other fairness violation notions

*Proof of Theorem 1.* We proceed to prove all the (in)equalities one by one:

- $0 \le I_S(\widehat{Y}; S|Z \in \mathcal{Z})$. This is well known and the proof can be found in any information theory textbook (Cover & Thomas, 1991).
- $I_1(\widehat{Y}; S|Z \in \mathcal{Z}) \le I_2(\widehat{Y}; S|Z \in \mathcal{Z})$. This is a known property of Rényi mutual information, but we provide a proof for completeness in Lemma 2.
- $I_2(\widehat{Y}; S|Z \in \mathcal{Z}) \le e^{I_2(\widehat{Y};S|Z\in\mathcal{Z})} - 1$. This follows from the fact that $x \le e^x - 1$.
- $e^{I_2(\widehat{Y};S)|Z\in\mathcal{Z}} - 1 = D_R(\widehat{Y}; S|Z \in \mathcal{Z})$. This follows from simple algebraic manipulation.

□

**Lemma 2.** *Let $\widehat{Y}, S, Z$ be discrete or continuous random variables. Then:*

*(a) For any $\alpha, \beta \in [1, \infty]$, $I_\beta(\widehat{Y}; S|Z \in \mathcal{Z}) \ge I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z})$ if $\beta > \alpha$.*

*(b) $\lim_{\alpha\to 1^+} I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z}) = I_1(\widehat{Y}; S) := \mathbb{E}_Z\left\{ D_{KL}(p_{\widehat{Y},S|Z}||p_{\widehat{Y}|Z} \otimes p_{S|Z})\Big| Z \in \mathcal{Z}\right\}$, where $I_1(\cdot;\cdot)$ denotes the Shannon mutual information and $D_{KL}$ is Kullback–Leibler divergence (relative entropy).*

*(c) For all $\alpha \in [1, \infty]$, $I_\alpha(\widehat{Y}; S|Z \in \mathcal{Z}) \ge 0$ with equality if and only if for all $z \in \mathcal{Z}$, $\widehat{Y}$ and $S$ are conditionally independent given $z$.*

*Proof.* *(a)* First assume $0 < \alpha < \beta < \infty$ and that $\alpha, \beta \ne 1$. Define $a = \alpha - 1$, and $b = \beta - 1$. Then the function $\phi(t) = t^{b/a}$ is convex for all $t \ge 0$, so by Jensen's inequality we have:

$$
\begin{aligned}
\frac{1}{b}\log\left(\mathbb{E}\left\{\left(\frac{p(\widehat{Y},S|Z)}{p(\widehat{Y}|Z)p(S|Z)}\right)^b \middle| Z \in \mathcal{Z}\right\}\right) &\ge \frac{1}{b}\log\left(\mathbb{E}\left\{\left(\frac{p(\widehat{Y},S|Z)}{p(\widehat{Y}|Z)p(S|Z)}\right)^a \middle| Z \in \mathcal{Z}\right\}^{b/a}\right) \\
&= \frac{1}{a}\log\left(\mathbb{E}\left\{\left(\frac{p(\widehat{Y},S|Z)}{p(\widehat{Y}|Z)p(S|Z)}\right)^a \middle| Z \in \mathcal{Z}\right\}\right).
\end{aligned} \tag{10}
$$

Now suppose $\alpha = 1$. Then by the monotonicity for $\alpha \ne 1$ proved above, we have $I_1(\widehat{Y}; S) = \lim_{\alpha\to 1^-} I_\alpha(\widehat{Y}; S) = \sup_{\alpha\in(0,1)} I_\alpha(\widehat{Y}; S) \le \inf_{\alpha>1} I_\alpha(\widehat{Y}; S)$. Also, $I_\infty(\widehat{Y}; S) = \lim_{\alpha\to\infty} I_\alpha(\widehat{Y}; S) = \sup_{\alpha>0} I_\alpha(\widehat{Y}; S)$.

*(b)* This is a standard property of the cumulant generating function (see (Dembo & Zeitouni, 2009)).

*(c)* It is straightforward to observe that independence implies that Rényi mutual information vanishes. On the other hand, if Rényi mutual information vanishes, then part (a) implies that Shannon mutual information also vanishes, which implies the desired conditional independence. □

*Proof of Theorem 2.* The proof is completed using the following pieces.

- $0 \le |\rho(\widehat{Y}, S|Z \in \mathcal{Z})| \le \rho_R(\widehat{Y}, S|Z \in \mathcal{Z})$. This is obvious from the definition of $\rho_R(\widehat{Y}, S|Z \in \mathcal{Z})$.
- $\rho_R(\widehat{Y}, S|Z \in \mathcal{Z}) \le D_R(\widehat{Y}; S|Z \in \mathcal{Z})$. This follows from Theorem 7.
- Notice that if $|\mathcal{S}| = 2$, Theorem 7 implies that $D_R(\widehat{Y}; S|Z \in \mathcal{Z}) = \rho_R(\widehat{Y}, S|Z \in \mathcal{Z})$.

□

**Theorem 7.** *Suppose that $\mathcal{S} = [k]$. Let the $k \times k$ matrix $P$ be defined as $P = \{P_{ij}\}_{i,j \in [k]\times[k]}$, where*

$$P_{ij} := \frac{1}{\sqrt{p_S(i)p_S(j)}} \int_{y \in \mathcal{Y}} \left( \frac{p_{\widehat{Y},S}(y,i)p_{\widehat{Y},S}(y,j)}{p_{\widehat{Y}}(y)} \right) dy. \tag{11}$$

*Let $1 = \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k \geq 0$ be the eigenvalues of $P$. Then,*

$$\rho_R(\widehat{Y}, S) = \sigma_2, \tag{12}$$

$$D_R(\widehat{Y}; S) = \mathrm{Tr}(P) - 1 = \sum_{i=2}^{k} \sigma_i. \tag{13}$$

*Proof.* Eq. (12) is proved in (Witsenhausen, 1975, Section 3). To prove Eq. (13), notice that

$$\begin{aligned}
\mathrm{Tr}(P) &= \sum_{i \in [k]} P_{ii} \\
&= \sum_{i \in [k]} \frac{1}{p_S(i)} \int_{y \in \mathcal{Y}} \left( \frac{p_{\widehat{Y},S}(y,i)^2}{p_{\widehat{Y}}(y)} \right) dy \\
&= E_{\widehat{Y},S} \left\{ \left( \frac{p_{\widehat{Y},S}(\widehat{Y},S)}{p_{\widehat{Y}}(\widehat{Y})p_S(S)} \right) \right\} \\
&= 1 + D_R(\widehat{Y}; S),
\end{aligned}$$

which completes the proof. □

*Proof of Theorem 3.* It suffices to prove the inequality for $L_1$, as $L_q$ is bounded above by $L_1$ for all $q \geq 1$. The proof for the case where $Z = 0$ and $\mathcal{Z} = \{0\}$ follows from the following set of inequalities:

$$L_1(\widehat{Y}, S | Z \in \mathcal{Z}) = \sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} \left| p_{\widehat{Y},S}(y,s) - p_{\widehat{Y}}(y)p_S(s) \right| dy \tag{14}$$

$$= \sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} \sqrt{p_{\widehat{Y}}(y)p_S(s)} \frac{\left| p_{\widehat{Y},S}(y,s) - p_{\widehat{Y}}(y)p_S(s) \right|}{\sqrt{p_{\widehat{Y}}(y)p_S(s)}} dy \tag{15}$$

$$\leq \sqrt{\left( \sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} p_{\widehat{Y}}(y)p_S(s) dy \right) \left( \sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} \left( \frac{(p_{\widehat{Y},S}(y,s) - p_{\widehat{Y}}(y)p_S(s))^2}{p_{\widehat{Y}}(y)p_S(s)} \right) \right)} \tag{16}$$

$$\leq \sqrt{\sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} \left( \frac{(p_{\widehat{Y},S}(y,s) - p_{\widehat{Y}}(y)p_S(s))^2}{p_{\widehat{Y}}(y)p_S(s)} \right) dy} \tag{17}$$

$$= \sqrt{D_R(\widehat{Y}; S)}, \tag{18}$$

where Eq. (16) follows from Cauchy-Schwarz inequality, and Eq. (18) follows from Lemma 3. The extension to general $Z$ and $\mathcal{Z}$ is immediate by observing that $\rho(\widehat{Y}, S | Z \in \mathcal{Z}) = \mathbb{E}_Z \left[ \rho(\widehat{Y}, S|Z) \middle| Z \in \mathcal{Z} \right]$, $\rho_R(\widehat{Y}, S | Z \in \mathcal{Z}) = \mathbb{E}_Z \left[ \rho_R(\widehat{Y}, S|Z) \middle| Z \in \mathcal{Z} \right]$, and $D_R(\widehat{Y}, S | Z \in \mathcal{Z}) = \mathbb{E}_Z \left[ D_R(\widehat{Y}, S|Z) \middle| Z \in \mathcal{Z} \right]$.

□

**Lemma 3.** *We have*

$$D_R(\widehat{Y}; S) = \sum_{s \in \mathcal{S}} \int_{y \in \mathcal{Y}} \left( \frac{(p_{\widehat{Y},S}(y,s) - p_{\widehat{Y}}(y)p_S(s))^2}{p_{\widehat{Y}}(y)p_S(s)} \right) dy. \tag{19}$$

*Proof.* The proof follows from the following set of identities:

$$\sum_{s\in\mathcal{S}}\int_{y\in\mathcal{Y}}\left(\frac{(p_{\widehat{Y},S}(y,s)-p_{\widehat{Y}}(y)p_S(s))^2}{p_{\widehat{Y}}(y)p_S(s)}\right)dy = \sum_{s\in\mathcal{S}}\int_{y\in\mathcal{Y}}\frac{(p_{\widehat{Y},S}(y,s))^2}{p_{\widehat{Y}}(y)p_S(s)}dy - 2\sum_{s\in\mathcal{S}}\int_{y\in\mathcal{Y}}p_{\widehat{Y},S}(y,s)dy + \sum_{s\in\mathcal{S}}\int_{y\in\mathcal{Y}}p_{\widehat{Y}}(y)p_S(s)dy \tag{20}$$

$$= E\left\{\frac{p_{\widehat{Y},S}(\widehat{Y},S)}{p_{\widehat{Y}}(\widehat{Y})p_S(S)}\right\}-1 \tag{21}$$

$$= D_R(\widehat{Y};S). \tag{22}$$

□

Next, we present some alternative fairness definitions and show that they are also upper bounded by ERMI.

**Definition 9** (conditional demographic parity $L_\infty$ violation)**.** *Given a predictor $\widehat{Y}$ supported on $\mathcal{Y}$ and a discrete sensitive attribute $S$ supported on a finite set $\mathcal{S}$, we define the conditional demographic parity violation by:*

$$\widetilde{dp}(\widehat{Y}|S) := \sup_{\widehat{y}\in\mathcal{Y}}\max_{s\in\mathcal{S}}\left|p_{\widehat{Y}|S}(\widehat{y}|s)-p_{\widehat{Y}}(\widehat{y})\right|. \tag{23}$$

First, we show that $\widetilde{\mathrm{dp}}(\widehat{Y}|S)$ is a reasonable notion of fairness violation.

**Lemma 4.** *$\widetilde{dp}(\widehat{Y}|S)=0$ iff (if and only if) $\widehat{Y}$ and $S$ are independent.*

*Proof.* By definition, $\widetilde{\mathrm{dp}}(\widehat{Y}|S)=0$ iff for all $\widehat{y}\in\mathcal{Y}, s\in\mathcal{S}$, $p_{\widehat{Y},S}(\widehat{y}|s)=p_{\widehat{Y}}(\widehat{y})$ iff $\widehat{Y}$ and $S$ are independent (since we always assume $p(s)>0$ for all $s\in\mathcal{S}$). □

**Theorem 8** (ERMI is stronger than conditional demographic parity $L_\infty$ violation)**.** *Let $\widehat{Y}$ be a discrete or continuous random variable supported on $\mathcal{Y}$, and $S$ be a discrete random variable supported on a finite set $\mathcal{S}$. Denote $p_S^{\min} := \min_{s\in\mathcal{S}} p_S(s) > 0$. Then,*

$$0 \le \widetilde{dp}(\widehat{Y}|S) \le \frac{1}{p_S^{\min}}\sqrt{D_R(\widehat{Y};S)}. \tag{24}$$

*Proof.* The proof follows from the following set of (in)equalities:

$$\left(\widetilde{\mathrm{dp}}(\widehat{Y}|S)\right)^2 = \sup_{\widehat{y}\in\mathcal{Y}}\max_{s\in\mathcal{S}}\left(p_{\widehat{Y}|S}(\widehat{y}|s)-p_{\widehat{Y}}(\widehat{y})\right)^2 \tag{25}$$

$$\le \frac{1}{(p_S^{\min})^2}\sup_{\widehat{y}\in\mathcal{Y}}\max_{s\in\mathcal{S}}\left(p_{\widehat{Y},S}(\widehat{y},s)-p_{\widehat{Y}}(\widehat{y})p_S(s))\right)^2 \tag{26}$$

$$\le \frac{1}{(p_S^{\min})^2}\int_{\widehat{y}\in\mathcal{Y}}\sum_{s\in\mathcal{S}}\left(p_{\widehat{Y},S}(\widehat{y},s)-p_{\widehat{Y}}(\widehat{y})p_S(s))\right)^2 \tag{27}$$

$$= \frac{1}{(p_S^{\min})^2}D_R(\widehat{Y};S), \tag{28}$$

where Eq. (28) follows from Theorem 3. □

**Definition 10** (conditional equal opportunity $L_\infty$ violation (Hardt et al., 2016))**.** *Let $Y, \widehat{Y}$ take values in $\mathcal{Y}$ and let $\mathcal{A}\subseteq\mathcal{Y}$ be a compact subset denoting the advantaged outcomes (For example, the decision "to interview" an individual or classify an individual as a "low risk" for financial purposes).*

*We define the conditional equal opportunity $L_\infty$ violation of $\widehat{Y}$ with respect to the sensitive attribute $S$ and the advantaged outcome $\mathcal{A}$ by*

$$\widetilde{eo}(\widehat{Y}|S, Y \in \mathcal{A}) := \mathbb{E}_Y \left\{ \sup_{\widehat{y}\in\mathcal{Y}} \max_{s\in\mathcal{S}} \left| p_{\widehat{Y},S|Y}(\widehat{y}|s, Y) - p_{\widehat{Y}|Y}(\widehat{y}|Y) \right| Y \in \mathcal{A} \right\}. \tag{29}$$

**Theorem 9** (ERMI is stronger than conditional equal opportunity $L_\infty$ violation)**.** *Let $\widehat{Y}$, $Y$, be discrete or continuous random variables supported on $\mathcal{Y}$, and let $S$ be a discrete random variable supported on a finite set $\mathcal{S}$. Let $\mathcal{A} \subseteq \mathcal{Y}$ be a compact subset of $\mathcal{Y}$.*

*Denote $p^{\min}_{S|\mathcal{A}} = \min_{s\in\mathcal{S}, y\in\mathcal{A}} p_{S|Y}(s|y)$. Then,*

$$0 \leq \widetilde{eo}(\widehat{Y}|S, Y \in \mathcal{A}) \leq \frac{1}{p^{\min}_{S|\mathcal{A}}} \sqrt{D_R(\widehat{Y}; S|Y \in \mathcal{A})}. \tag{30}$$

*Proof.* Notice that the same proof for Theorem 8 would give that for all $y \in \mathcal{A}$:

$$\begin{aligned}
0 \leq \sup_{\widehat{y}\in\mathcal{Y}} \max_{s\in\mathcal{S}} \left| p_{\widehat{Y},S|Y}(\widehat{y}|s, y) - p_{\widehat{Y}|Y}(\widehat{y}|y) \right| &:= \widetilde{\text{eo}}(\widehat{Y}|S, Y = y) \\
&\leq \frac{1}{p^{\min}_{S|y}(y)} \sqrt{D_R(\widehat{Y}; S|Y = y)} \\
&\leq \frac{1}{p^{\min}_{S|\mathcal{C}}} \sqrt{D_R(\widehat{Y}; S|Y = y)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\widetilde{\text{eo}}(\widehat{Y}|S, Y \in \mathcal{A}) &= \mathbb{E}_Y \left\{ \widetilde{\text{eo}}(\widehat{Y}|S, Y) \middle| Y \in \mathcal{A} \right\} \\
&\leq \frac{1}{p^{\min}_{S|\mathcal{A}}} \mathbb{E}_Y \left\{ \sqrt{D_R(\widehat{Y}; S|Y)} \middle| Y \in \mathcal{A} \right\} \\
&\leq \frac{1}{p^{\min}_{S|\mathcal{A}}} \sqrt{\mathbb{E}_Y \left\{ D_R(\widehat{Y}; S|Y) \middle| Y \in \mathcal{A} \right\}} \\
&= \frac{1}{p^{\min}_{S|\mathcal{A}}} \sqrt{D_R(\widehat{Y}; S|Y \in \mathcal{A})},
\end{aligned}$$

where the last inequality follows from Jensen's inequality. This completes the proof. ☐

## D FERMI: objective and algorithm

*Proof of Theorem 4.* Let $W^* \in \arg\max_{W \in \mathbb{R}^{k\times m}} -\operatorname{Tr}(WP_{\widehat{y}}W^T) + 2\operatorname{Tr}(WP_{\widehat{y},s}P_s^{-1/2})$. We will compute $W^*$ and plug it in the RHS of Eq. (4) to show the equality in Eq. (4). Setting the derivative of the expression on the RHS equal to zero leads to:

$$-2WP_{\widehat{y}} + 2P_s^{-1/2}P_{\widehat{y},s}^T = 0 \implies W^* = P_{\widehat{y}}^{-1}P_{\widehat{y},s}^T P_s^{-1/2}.$$

Plugging this expression for $W^*$, we have

$$\begin{aligned}
&\max_{W\in\mathbb{R}^{k\times m}} -\operatorname{Tr}(WP_{\widehat{y}}W^T) + 2\operatorname{Tr}(WP_{\widehat{y},s}P_s^{-1/2}) \\
&\quad= -\operatorname{Tr}(P_s^{-1/2}P_{\widehat{y},s}^T P_{\widehat{y}}^{-1}P_{\widehat{y}}P_{\widehat{y}}^{-1}P_s^{-1/2}) + 2\operatorname{Tr}(P_s^{-1/2}P_{\widehat{y},s}^T P_{\widehat{y}}^{-1}P_{\widehat{y}}P_{\widehat{y}}^{-1}P_s^{-1/2}) \\
&\quad= \operatorname{Tr}(P_s^{-1/2}P_{\widehat{y},s}^T P_{\widehat{y}}^{-1}P_{\widehat{y},s}P_s^{-1/2}) \\
&\quad= \operatorname{Tr}(P_s^{-1}P_{\widehat{y},s}^T P_{\widehat{y}}^{-1}P_{\widehat{y},s}).
\end{aligned}$$

Writing out the matrix multiplication explicitly in the last expression, we have

$$P_s^{-1}P_{\widehat{y},s}^T P_{\widehat{y}}^{-1}P_{\widehat{y},s} = UV^T,$$

where $U_{i,j} = \widehat{p}_S(i)^{-1}\widehat{p}_{\widehat{Y},S}(j,i)$ and $V_{i,j} = \hat{p}_{\widehat{Y}}(j)^{-1}\hat{p}_{\widehat{Y},S}(j,i)$, for $i \in [k], j \in [m]$. Hence

$$\begin{aligned}
\max_{W\in\mathbb{R}^{k\times m}} -\operatorname{Tr}(WP_{\widehat{y}}W^T) + 2\operatorname{Tr}(WP_{\widehat{y},s}P_s^{-1/2}) &= \operatorname{Tr}(UV^T) \\
&= \sum_{i\in[k]}\sum_{j\in[m]} \frac{p_{\widehat{Y},S}(j,i)^2}{p_S(i)p_{\widehat{Y}}(j)} \\
&= D_R(\widehat{Y}; S),
\end{aligned}$$

which completes the proof. □

Next, we move to the statement and proof of the precise version of Theorem 5. We first recall some basic definitions:

**Definition 11.** *A function $f$ is $\beta$-smooth if for all $\mathbf{u}, \mathbf{u}'$, we have $\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{u})\| \le \beta\|\mathbf{u} - \mathbf{u}'\|$.*

**Definition 12.** *A point $\boldsymbol{\theta}$ is an $\epsilon$-stationary point of a differentiable function $\Phi$ if $\|\nabla\Phi(\boldsymbol{\theta})\| \le \epsilon$.*

**Assumption 1.**

- *$\ell$ is twice differentiable, $L_\ell$-Lipscthiz, and $\beta_\ell$-smooth in $\boldsymbol{\theta}$.*
- $\|\nabla_\theta P_{\hat{y}}\|_2 := \|\nabla_\theta \operatorname{vec}(P_{\widehat{y}})\|_2 \le L_y$ *and* $\max_{l\in[m]} \|\nabla_\theta\left((P_{\widehat{y}})_{l,l}\right)\|_2 \le \widetilde{L}_y$
- $\max_{l\in[m]} \|\nabla^2_{\theta\theta}(P_{\widehat{y}})_{l,l}\|_2 \le \beta_y$.
- $\|\nabla_\theta P_{\widehat{y},s}^T\|_2 := \|\nabla_\theta \operatorname{vec}(P_{\widehat{y},s}^T)\|_2 \le L_{ys}$ *and* $\max_{l\in[m],j\in[k]} \|\nabla_\theta\left((P_{\widehat{y},s})_{l,m}\right)\|_2 \le \widetilde{L}_{ys}$
- $\max_{l\in[m],j\in[k]} \|\nabla^2_{\theta\theta}(P_{\widehat{y},s})_{l,j}\|_2 \le \beta_{y,s}$.

**Theorem 10** (Precise statement of Theorem 5)**.** *Denote*

$$f(\boldsymbol{\theta}, W) = \frac{1}{N}\sum_{i\in[N]} \ell\big(\mathbf{x}_i, y_i; \boldsymbol{\theta}\big) + \lambda\left(-\operatorname{Tr}(WP_{\widehat{y}}W^T) + 2\operatorname{Tr}(WP_{\widehat{y},s}P_s^{-1/2}) - 1\right).$$

*Set $\mathcal{W} := B_F(0, 2D) \subset \mathbb{R}^{k\times m}$ (Frobenius norm ball of radius $2D$), $D := \frac{\sqrt{mk}}{\hat{p}_{\widehat{y}}^{\min}\sqrt{\hat{p}_s^{\min}}}$. Denote $\Delta_\Phi := \Phi(\theta_0) - \min_\theta \Phi(\theta)$, where $\Phi(\theta) := \max_{W\in\mathcal{W}} f(\boldsymbol{\theta}, W)$. In Algorithm 1, choose the step-sizes as $\eta_\theta = \Theta(1/\kappa^2\beta)$ and $\eta_W = \Theta(1/\beta)$ and mini-batch size as $M = \Theta\left(\max\left\{1, \kappa\sigma^2\epsilon^{-2}\right\}\right)$. Then under Assumption 1, the iteration complexity of Algorithm 1 to return an $\epsilon$-stationary point of $f$ is bounded by*

$$\mathcal{O}\left(\frac{\kappa^2\beta\Delta_\Phi + \kappa\beta^2 D^2}{\epsilon^2}\right),$$

*which gives the total stochastic gradient complexity of*

$$\mathcal{O}\left(\left(\frac{\kappa^2\beta\Delta_\Phi + \kappa\beta^2 D^2}{\epsilon^2}\right)\max\left\{1, \kappa\sigma^2\epsilon^{-2}\right\}\right),$$

*where*

$$\begin{aligned}
\beta &= \beta_l + 8\lambda D^2\beta_y + 4\lambda\frac{1}{\sqrt{\hat{p}_s^{\min}}}\left(\sqrt{m}k^{3/2}D\beta_{ys}\right) + 2\lambda + 4\lambda\left(DL_y + \frac{L_{ys}}{\sqrt{\hat{p}_s^{\min}}}\right), \\
\mu &= 2\lambda\hat{p}_{\widehat{y}}^{\min}, \\
\kappa &= \beta/\mu, \\
\sigma^2 &= 2\left(L_\ell + 2\lambda\widetilde{L}_y D^2 + 4\lambda\frac{D}{\sqrt{\hat{p}_s^{\min}}}\sqrt{mk}\widetilde{L}_{ys}\right)^2 + 2\left(2\lambda D + 2(\hat{p}_s^{\min})^{-1/2}\sqrt{mk}\right)^2.
\end{aligned}$$

The theorem follows from Theorem 4.5 in (Lin et al., 2020) combined with the following technical lemmas. We assume Assumption 1 holds for the remainder of the proof of Theorem 10:

**Lemma 5.** *Let*

$$\begin{aligned}
f(\boldsymbol{\theta}, W) &= \frac{1}{N}\sum_{i\in[N]}\ell\big(\mathbf{x}_i, y_i; \boldsymbol{\theta}\big) + \lambda\left(-\operatorname{Tr}(WP_{\widehat{y}}W^T) + 2\operatorname{Tr}(WP_{\widehat{y},s}P_s^{-1/2}) - 1\right) \\
&:= \frac{1}{N}\sum_{i\in[N]} g(\boldsymbol{\theta}, W, \mathbf{x}_i, y_i).
\end{aligned}$$

*Then*

1. *$f$ is $\beta$-smooth, where $\beta = \beta_l + 8\lambda D^2\beta_y + 4\lambda\frac{1}{\sqrt{\hat{p}_s^{\min}}}\left(\sqrt{m}k^{3/2}D\beta_{ys}\right) + 2\lambda + 4\lambda\left(DL_y + \frac{L_{ys}}{\sqrt{\hat{p}_s^{\min}}}\right)$.*

2. *$f(\boldsymbol{\theta}, \cdot)$ is $2\lambda\hat{p}_{\widehat{y}}^{\min}$-strongly concave for all $\boldsymbol{\theta}$.*

3. *$\|W^*\|_F \leq D$, where $D$ is as defined in Theorem 10 and $W^* \in \arg\max_{W\in\mathbb{R}^{k\times m}}$ denotes any maximizer of $f(\boldsymbol{\theta}, W)$.*

*Proof.* By Assumption 1, $g$ is twice continuously differentiable. Hence for part 1, it suffices to upper bound the spectral norm of the second derivative of $g(\cdot,\cdot,\mathbf{z})$ by $\beta$ for all $\mathbf{z} = (\mathbf{x}, y)$, where we vectorize and then differentiate with respect to $w := \operatorname{vec} W$ and/or $\boldsymbol{\theta}$, so that the resulting first and second derivatives are always vectors or a matrices (not tensors). Notice that $g(\boldsymbol{\theta}, w, \mathbf{z}) = \ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda w^T(P_{\widehat{y}} \otimes \mathbf{I})w + 2\lambda(\operatorname{vec}(W))^T P_{\widehat{y},s}P_s^{-1/2} - \lambda$ and

$$\nabla^2 g(\boldsymbol{\theta}, w, \mathbf{z}) = \begin{pmatrix} \nabla^2_{\theta\theta} g(\boldsymbol{\theta}, w, \mathbf{z}) & \nabla^2_{\theta w} g(\boldsymbol{\theta}, w, \mathbf{z}) \\ \nabla^2_{w\theta} g(\boldsymbol{\theta}, w, \mathbf{z}) & \nabla^2_{ww} g(\boldsymbol{\theta}, w, \mathbf{z}) \end{pmatrix}.$$

Further, by the definition of operator norm, we have

$$\|\nabla^2 g(\boldsymbol{\theta}, w, \mathbf{z})\|_2 \leq \|\nabla^2_{\theta\theta} g(\boldsymbol{\theta}, w, \mathbf{z})\|_2 + 2\|\nabla^2_{\theta w} g(\boldsymbol{\theta}, w, \mathbf{z})\|_2 + \|\nabla^2_{ww} g(\boldsymbol{\theta}, w, \mathbf{z})\|_2.$$

Now we vectorize all matrices and then compute derivatives of $g$ with respect to $\theta$ and $\operatorname{vec}(W)$:

$$\nabla_\theta g(\boldsymbol{\theta}, w, \mathbf{z}) = \nabla_\theta \ell(\mathbf{z}, \boldsymbol{\theta}) - 2\lambda \nabla_\theta \operatorname{vec}(P_{\widehat{y}})^T \operatorname{vec}(W^T W) + 2\lambda \nabla_\theta \operatorname{vec}(P_{\widehat{y},s})^T \operatorname{vec}(W^T P_s^{-1/2}) \tag{31}$$

$$= \nabla_\theta \ell(\mathbf{z}, \boldsymbol{\theta}) - 2\lambda \left[ \sum_{l \in [m], i \in [k]} W_{i,l}^2 \nabla_\theta \left( (P_{\widehat{y}})_{l,l} \right) \right] + 2\lambda \left[ \sum_{j \in [m], i \in [k]} W_{i,j} (\nabla_\theta \left( P_{\widehat{y}s} \right)_{j,i}) \, (P_s^{-1/2})_{i,i} \right]; \tag{32}$$

$$\nabla_w g(\boldsymbol{\theta}, w, \mathbf{z}) = -2\lambda W P_{\widehat{y}} + 2\lambda P_s^{-1/2} P_{\widehat{y},s}^T. \tag{33}$$

Differentiating again yields:

$$\begin{aligned}
\nabla_{ww}^2 g(\boldsymbol{\theta}, w, \mathbf{z}) &= -2\lambda P_{\widehat{y}} \otimes \mathbf{I}_k; \\
\nabla_{w\theta}^2 g(\boldsymbol{\theta}, w, \mathbf{z}) &= \frac{\partial}{\partial \theta} \frac{\partial g(\boldsymbol{\theta}, w, \mathbf{z})}{\partial w} = -2\lambda (\mathbf{I}_m \otimes W) \nabla_\theta P_{\widehat{y}} + 2\lambda (\mathbf{I}_m \otimes P_s^{-1/2}) \nabla_\theta \operatorname{vec}(P_{\widehat{y},s}^T); \\
\nabla_{\theta\theta}^2 g(\boldsymbol{\theta}, w, \mathbf{z}) &= \nabla_\theta^2 \ell(\mathbf{z}, \boldsymbol{\theta}) - 2\lambda \left[ \sum_{l \in [m], i \in [k]} W_{i,l}^2 \nabla_{\theta\theta}^2 \left( (P_{\widehat{y}})_{l,l} \right) \right] \\
&\quad + 2\lambda \left[ \sum_{j \in [m], i \in [k]} W_{i,j} (\nabla_{\theta\theta}^2 \left( P_{\widehat{y}s} \right)_{j,i}) \, (P_s^{-1/2})_{i,i} \right].
\end{aligned}$$

Then to establish part 1, use Assumption 1, Clairaut's theorem, the definitions of the matrices and fact that their entries are in $[0, 1]$, the relations $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ and $\|\operatorname{vec} W\|_1 \leq \sqrt{mk} \|\operatorname{vec} W\|_2 = \sqrt{mk} \|W\|_F$, and the fact that $\|A \otimes B\|_2 = \|A\|_2 \|B\|_2$ to bound the spectral norm of each second derivative above.

The strong concavity statement follows by noticing $\nabla_{ww}^2 g(\boldsymbol{\theta}, W) \preccurlyeq -\mu \mathbf{I}$ iff $P_{\widehat{y}} \succcurlyeq \frac{\mu}{2\lambda} \mathbf{I}$ iff $\min_{i \in [m]} p_{\widehat{y}}(i) \geq \frac{\mu}{2\lambda}$.

Part 3 follows from the expression for $W^*$ in the proof of Theorem 4. □

**Lemma 6.** *Consider $f$ and $g$ as defined above. Then we have*

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}} \nabla g(\boldsymbol{\theta}, W, \mathbf{z}) &= \nabla f(\boldsymbol{\theta}, W), \\
\mathbb{E}_{\mathbf{z}} \|\nabla g(\boldsymbol{\theta}, W, \mathbf{z}) - \nabla f(\boldsymbol{\theta}, W)\|_2^2 &\leq 2 \left( L_\ell + 2\lambda \widetilde{L}_y D^2 + 4\lambda \frac{D}{\sqrt{\hat{p}_s^{\min}}} \sqrt{mk} \widetilde{L}_{ys} \right)^2 \\
&\quad + 2 \left( 2\lambda D + 2 (\hat{p}_s^{\min})^{-1/2} \sqrt{mk} \right)^2,
\end{aligned}$$

*where both expectations are with respect to the empirical distribution on $\{\mathbf{z}_i\}_{i \in [N]}$.*

*Proof.* The first statement is obvious. The second follows from Eq. (32) in the proof of Lemma 5, since

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{z}}\|\nabla g(\boldsymbol{\theta}, W, \mathbf{z}) - \nabla f(\boldsymbol{\theta}, W)\|_2^2 \\
&\quad = \frac{1}{N}\sum_{i=1}^{N}\|\nabla g(\boldsymbol{\theta}, W, z_i)\|_2^2 - \frac{1}{N^2}\sum_{i,j=1}^{N}\langle \nabla g(\boldsymbol{\theta}, W, \mathbf{z}_i), \nabla g(\boldsymbol{\theta}, W, \mathbf{z}_j)\rangle \\
&\quad \le 2\sup_{\mathbf{z}_i}\|\nabla g(\boldsymbol{\theta}, W, \mathbf{z}_i)\|_2^2 \\
&\quad \le 2\sup_{\mathbf{z}}\left\{\|\nabla_\theta g(\boldsymbol{\theta}, W, \mathbf{z})\|^2 + \|\nabla_w g(\boldsymbol{\theta}, W, \mathbf{z})\|^2\right\} \\
&\quad \le 2\sup_{\mathbf{z}}\Bigg\{\Bigg\|\nabla_\theta \ell(\mathbf{z}, \boldsymbol{\theta}) - 2\lambda\left[\sum_{l\in[m], i\in[k]} W_{i,l}^2 \nabla_\theta\left((P_{\widehat{y}})_{l,l}\right)\right] \\
&\qquad\qquad + 2\lambda\left[\sum_{j\in[m], i\in[k]} W_{i,j}(\nabla_\theta\left(P_{\widehat{y}s}\right)_{j,i})\,(P_s^{-1/2})_{i,i}\right]\Bigg\|_2^2\Bigg\} \\
&\qquad + 2\left\|-2\lambda W P_{\widehat{y}} + 2\lambda P_s^{-1/2} P_{\widehat{y},s}^T\right\|_2^2.
\end{aligned}
$$

Then use Assumption 1 and basic norm inequalities to bound the norm of each term. □

# E Experiment details & additional results

## E.1 Model description

For all the experiments, the model's output is of the form $O = \mathrm{softmax}(Wx + b)$. The model outputs are treated as conditional probabilities $\mathbf{p}(\widehat{y} = i|x) = O_i$ which are then used to estimate the ERMI regularizer. We encode the true class label $Y$ and sensitive attribute $S$ using one-hot encoding. We define $\ell(\cdot)$ as the cross-entropy measure between the one-hot encoded class label $Y$ and the predicted output vector $O$.

We use logistic regression as the base classification model for all experiments in Fig. 1. The choice of logistic regression is due to the fact that all of the existing approaches demonstrated in Fig. 1, use the same classification model. The model parameters are estimated using the algorithm described in Algorithm 1. The trade-off curves for FERMI are generated by sweeping across different values for $\lambda \in [0, 10000]$. The learning rates $\eta_\theta, \eta_w$ is constant during the optimization process and is chosen from the interval $[0.0005, 0.01]$ for all datasets. Moreover, the number of iterations $T$ for experiments in Fig. 1 is fixed to $2000$. Since the training and test data for the Adult dataset are separated and fixed, we do not consider confidence intervals for the test accuracy. We generate ten distinct train/test sets for each one of the German and COMPAS datasets by randomly sampling $80\%$ of data points as the training data and the rest $20\%$ as the test data. For a given method in Fig. 1, the corresponding curve is generated by taking the average test accuracy on $10$ training/test datasets. Furthermore, the confidence intervals are estimated based on the test accuracy's standard deviation on these $10$ datasets.

To perform the experiments in Sec. 5.2 we use a a linear model with softmax activation. The model parameters are estimated using the algorithm described in Sec. 5. The data set is cleaned and processed as described in (Kearns et al., 2018). The trade-off curves for FERMI are generated by sweeping across different values for $\lambda$ in $[0, 100]$ interval, learning rate $\eta$ in $[0.0005, 0.01]$, and number of iterations $T$ in $[50, 200]$. The data set is cleaned and processed as described in (Kearns et al., 2018).

For the experiments in Sec. 5.3, we create the synthetic color MNIST as described by Li & Vasconcelos (2019). We set the value $\sigma = 0$. In Fig. 3, we compare the performance of stochastic solver (Algorithm 1) against the baselines. We use a mini-batch of size 512 when using the stochastic solver. The color MNIST data has 60000 training samples, so using the stochastic solver gives a speedup of around 100x for each iteration, and an overall speedup of around 40x. We present our results on two neural network architectures; namely, LeNet-5 (Lecun et al., 1998) and a Multi-layer perceptron (MLP). We set the MLP with two hidden layers (with 300 and 100 nodes) and an output layer with ten nodes. A ReLU activation follows each hidden layer, and a softmax activation follows the output layer.

Some general advice for tuning $\lambda$: Larger value for $\lambda$ generally translates to better fairness, but one must be careful to not use a very large value for $\lambda$ as it could lead to poor generalization performance of the model. The optimal values for $\lambda$, $\eta$, and $T$ largely depend on the data and intended application. We recommend starting with $\lambda \approx 10$. In Appendix E.4, we can observe the effect of changing $\lambda$ on the model accuracy and fairness for the COMPAS dataset.

## E.2 More comparison to (Mary et al., 2019)

The algorithm proposed by Mary et al. (2019) backpropagates the batch estimate of ERMI, which is biased especially for small minibatches. Our work uses a correct and unbiased implementation of a stochastic ERMI estimator; Furthermore, they do not establish any convergence guarantees, and in fact their algorithm does not converge. See Fig. 4 for the evolution of *training loss* and *test accuracy* on setup of Table 1 in (Mary et al., 2019).

## E.3 Performance in the presence of outliers & class-imbalance

We also performed an additional experiment on Adult (setup of Fig 1) with a random 10% of sensitive attributes in *training* forced to 0. FERMI offers the most favorable tradeoffs on *clean test* data, however, all methods reach a higher plateau (see Fig 5). The interplay between fairness, robustness, and generalization is an important future direction. With respect to imbalanced sensitive groups, the

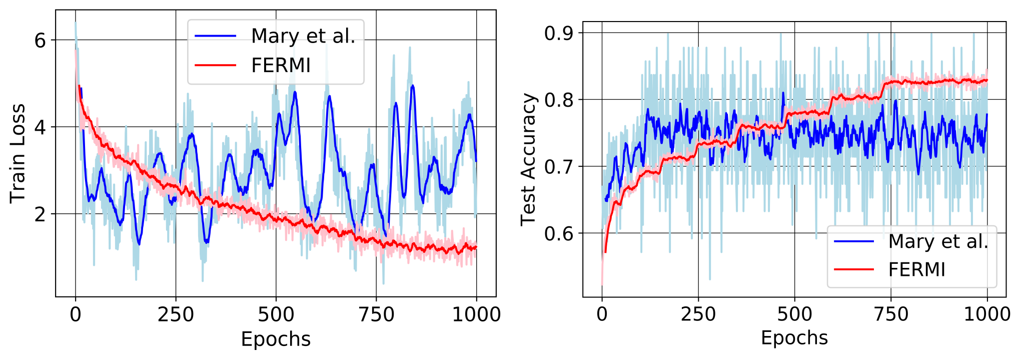


Figure 4: (Mary et al., 2019) fails to converge to a stationary point whereas our stochastic estimator easily converges

experiments in Fig 2 are on a naturally imbalanced dataset, where $\max_{s \in \mathcal{S}} p(s)/\min_{s \in \mathcal{S}} p(s) > 100$ for 3-18 sensitive attrib, and FERMI offers the favorable tradeoffs.

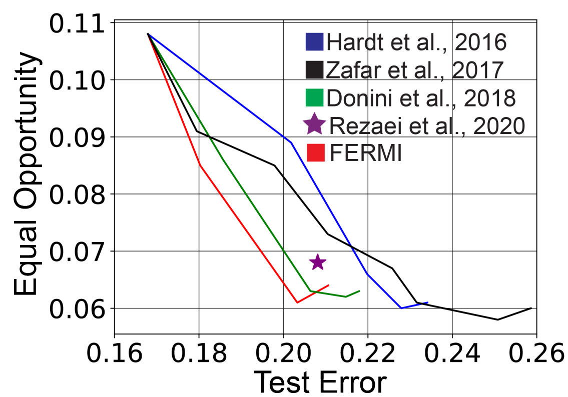


Figure 5: Comparing FERMI with other methods in the presence of outliers (random 10% of sensitive attributes in *training* forced to 0. FERMI achieves better trade-off.

## E.4 Effect of hyperparameter $\lambda$ on the accuracy-fairness tradeoffs

We run ERMI algorithm for the binary case to COMPAS dataset to investigate the effect of hyper-parameter tuning on the accuracy-fairness trade-off of the algorithm. As it can be observed in Fig. 6, by increasing $\lambda$ from 0 to 1000, test error (left axis, red curves) is slightly increased. On the other hand, the fairness violation (right axis, green curves) is decreased as we increase $\lambda$ to 1000. Moreover, for both notions of fairness (demographic parity with the solid curves and equality of opportunity with the dashed curves) the trade-off between test error and fairness follows the similar pattern. To measure the fairness violation, we use demographic parity violation and equality of opportunity violation defined in Section equation 5 for the solid and dashed curves respectively.

## E.5 Complete version of Figure 1 (with pre-processing and post-processing baselines)

In Figure 1 we compared FERMI with several state-of-the-art in-processing approaches. In the next three following figures we compare the in-processing approaches depicted in Figure 1 with pre-processing and post-processing methods including (Hardt et al., 2016; Kamiran et al., 2010; Feldman et al., 2015).

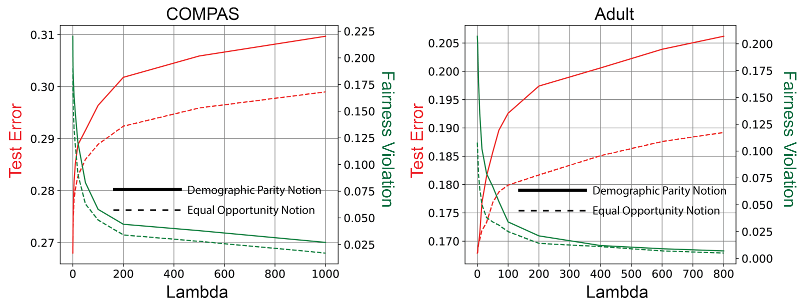


Figure 6: Tradeoff of fairness violation vs test error for FERMI algorithm on COMPAS and Adult datasets. The solid and dashed curves correspond to FERMI algorithm under the demographic parity and equality of opportunity notions accordingly. The left axis demonstrates the effect of changing $\lambda$ on the test error (red curves), while the right axis shows how the fairness of the model (measured by equality of opportunity or demographic parity violations) depends on changing $\lambda$.

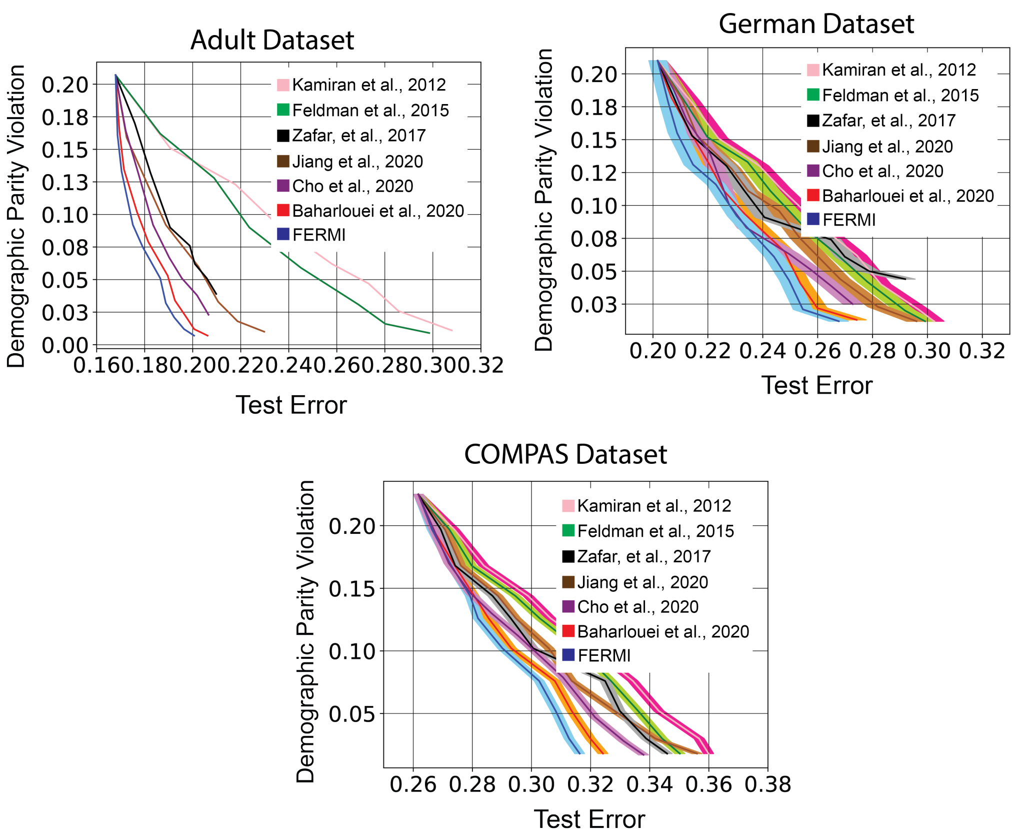


Figure 7: Tradeoff of demographic parity violation vs test error for FERMI algorithm on COMPAS, German, and Adult datasets.

## E.6 Description of datasets

All of the following datasets are publicly available at UCI repository.

**German Credit Dataset.**[3] German Credit dataset consists of 20 features (13 categorical and 7 numerical) regarding to social, and economic status of 1000 customers. The assigned task is to classify customers as good or bad credit risks. Without imposing fairness, the DP violation of the

[3] `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

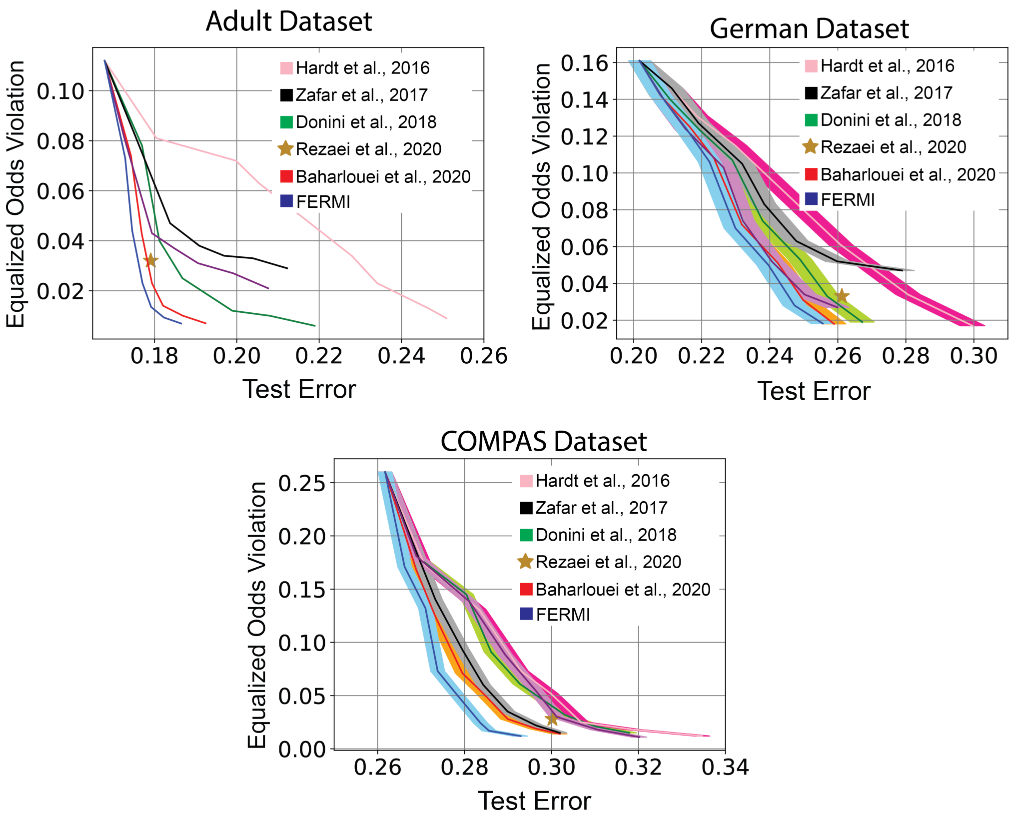


Figure 8: Tradeoff of equalized odds violation vs test error for FERMI algorithm on COMPAS, German, and Adult datasets.

trained model is larger than 20%. We choose 80% of customers as the train data and the remaining 20% customers as the test data. The sensitive attributes are gender, and marital-status.

**Adult Dataset.**[4] Adult dataset contains the census information of individuals including education, gender, and capital gain. The assigned classification task is to predict whether a person earns over 50k annually. The train and test sets are two separated files consisting of $32,000$ and $16,000$ samples respectively. We consider gender and race as the sensitive attributes (For the experiments involving one sensitive attribute, we have chosen gender). Learning a logistic regression model on the training dataset (without imposing fairness) shows that only 3 features out of 14 have larger weights than the gender attribute. Note that removing the sensitive attribute (gender), and retraining the model does not eliminate the bias of the classifier. the optimal logistic regression classifier in this case is still highly biased. For the clustering task, we have chosen 5 continuous features (Capital-gain, age, fnlwgt, capital-loss, hours-per-week), and $10,000$ samples to cluster. The sensitive attribute of each individual is gender.

**Communities and Crime Dataset**.[5] The dataset is cleaned and processed as described in (Kearns et al., 2018). Briefly, each record in this dataset summarizes aggregate socioeconomic information about both the citizens and police force in a particular U.S. community, and the problem is to predict whether the community has a high rate of violent crime.

**COMPAS Dataset**.[6] Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a famous algorithm which is widely used by judges for the estimation of likelihood of reoffending crimes. It is observed that the algorithm is highly biased against the black defendants.

[4] `https://archive.ics.uci.edu/ml/datasets/adult.`

[5] `http://archive.ics.uci.edu/ml/datasets/communities+and+crime`

[6] `https://www.kaggle.com/danofer/compass`

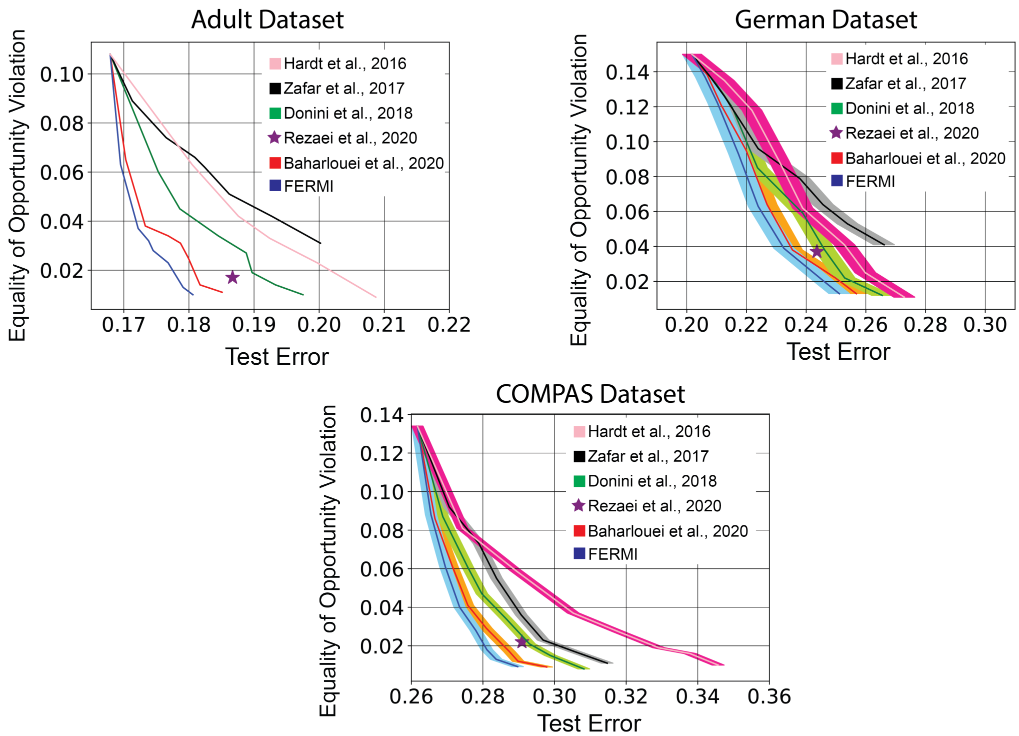


Figure 9: Tradeoff of equality of opportunity violation vs test error for FERMI algorithm on COMPAS, German, and Adult datasets.

The dataset contains features used by COMPAS algorithm alongside with the assigned score by the algorithm within two years of the decision.

**Colored MNIST Dataset.**[7] We use the code by Li & Vasconcelos (2019) to create a Colored MNIST dataset with $\sigma = 0$. We use the provided LeNet-5 model trained on the colored dataset for all baseline models of Baharlouei et al. (2020); Mary et al. (2019); Cho et al. (2020b) and FERMI, where we further apply the corresponding regularizer in the training process.

[7] `https://github.com/JerryYLi/Dataset-REPAIR/`

# F Anonymized code for experiments

The anonymized code for all of the experiments in this paper is available on Dropbox: https://www.dropbox.com/sh/516cm8olq0idpsd/AADD0LOcPWpx4AAhzsEkFTOca?dl=0