Jointly Robust Fairness: Overcoming Simultaneous Label and Attribute Noise

Gaurav Jain
Independent Researcher
Gurugram, India
jaingaurav101994@gmail.com

Abstract

Fairness in machine learning typically relies on the assumption of clean training data. However, in high-stakes applications like healthcare and criminal justice, privacy mechanisms and proxy variables frequently introduce simultaneous noise into both sensitive attributes and target labels. While existing methods address label or attribute noise in isolation, they fail under this "double noise" regime, often inadvertently amplifying bias. In this paper, we propose **Jointly Robust Fairness (JRF)**[1], a unified framework for fair learning under simultaneous data corruption. JRF integrates a Forward Loss correction to handle label noise with a novel Matrix-Inverse Robust Loss (MIRL) that algebraically recovers the true demographic parity gap from noisy attribute observations. We provide rigorous theoretical guarantees for our estimator, including a finite-sample concentration bound demonstrating that the sample complexity scales quadratically with the inverse of the noise transition determinant. Furthermore, the correction extends in closed form to asymmetric noise and to multi-class sensitive attributes. Extensive experiments on the Adult, Bank, and COMPAS datasets demonstrate that, in high-noise regimes ($\rho \geq 0.3$), our method reduces fairness violations by up to 91.8% relative to naive fair training, while a standard reweighting baseline amplifies bias.

1 Introduction

Machine learning models are increasingly deployed in high-stakes domains such as credit scoring, hiring, and criminal justice, where fairness is a critical requirement, highlighted by high-profile failures in recidivism prediction [1] and facial recognition [2]. A fundamental challenge in deploying these models is the quality of the training data. Standard fairness algorithms, such as those proposed by Hardt et al. [3] and Zafar et al. [4], assume access to clean ground-truth values for both the target label Y and the sensitive attribute A. However, in real-world applications, this assumption rarely holds.

Sources of Simultaneous Noise. In practice, noise stems from multiple distinct sources, often affecting both labels and attributes simultaneously. *Privacy Mechanisms:* Techniques like Differential Privacy [5] intentionally inject noise into sensitive attributes via Randomized Response [6] to protect user anonymity. *Proxy Labels:* In domains like healthcare, the label Y is often a proxy (e.g., "patient took medication") rather than the ground truth ("patient was cured"), introducing systematic label noise [7]. *Inferred Demographics:* In marketing and finance, race and gender are often inferred from names or geolocation using auxiliary models, introducing significant attribute noise [8]. When these factors combine, practitioners face a "double noise" regime where neither Y nor A is reliable.

Limitations of Prior Work. While robust learning is a well-studied field, existing methods largely treat these problems in isolation. For instance, loss-correction methods [9] handle noisy labels (Y)

[1] Code and implementation are available at `https://github.com/gauravjain10/JRF`.

but do not enforce fairness. Conversely, Lamy et al. [10], Wang et al. [11], and Celis et al. [12] addressed fairness when sensitive attributes (A) are noisy or inferred, while assuming clean labels. However, **no existing method addresses the joint scenario** where both Y and A are corrupted simultaneously without access to a clean validation set. Applying single-variable corrections in this setting is insufficient; as we show in our analysis, standard fairness regularizers can inadvertently amplify bias when the noise structure is complex.

Our Contribution. We introduce **Jointly Robust Fairness (JRF)**, a unified framework to handle simultaneous label and attribute noise. The contributions of the paper are three-fold:

1. We theoretically quantify how double noise corrupts the Demographic Parity metric, deriving a precise relationship between the observed and true fairness gaps.
2. We propose the *Matrix-Inverse Robust Loss (MIRL)*, a loss function that uses a noise transition matrix to recover the true group statistics during training.
3. We empirically demonstrate on the Adult, Bank, and COMPAS datasets that our method reduces fairness violations by up to **91.8%** in high-noise regimes ($\rho \geq 0.3$) relative to naive fair training, with accuracy within 1.5 percentage points of naive fair training on Adult ($\rho \leq 0.3$) and Bank.

2 Motivation: The Double-Blind Spot

Why is the joint noise scenario critical? Consider a healthcare risk prediction model. **Label Noise (Y):** The true target ("Heart Attack") is unobserved; the model relies on a noisy proxy ("Cardiologist Visit"), introducing false positives and negatives. **Attribute Noise (A):** For privacy, patient demographics are perturbed via Local Differential Privacy (LDP) before analysis. **Failure of Current Methods:** Standard fairness methods assume both the proxy label and LDP-perturbed race are ground truth. Consequently, they enforce equality on a biased proxy across statistically blended groups. This creates a "blind spot" where models report a near-zero fairness gap (formally defined in Section 4) on training data while remaining unfair in deployment. Our method, **JRF**, explicitly models this dual corruption to optimize for the unobserved reality.

3 Related Work

Fairness in Machine Learning. A vast body of work has focused on algorithmic fairness, primarily defining constraints such as Demographic Parity [13] and Equalized Odds [3]. Early methods focused on pre-processing data [14] or post-processing predictions [15]. More recent approaches integrate fairness constraints directly into the optimization objective via dual decomposition or regularization [4, 16]. However, most of these methods assume that both the labels and the sensitive attributes in the training data are ground truth.

Learning with Noisy Labels. Robust learning under label noise is a well-studied problem. Common techniques include loss correction [9], sample selection [17], and label smoothing [18, 19]. While effective for standard classification, these methods do not account for fairness. Imposing fairness constraints that ignore group-dependent label noise can reduce both accuracy and fairness [20].

Fairness under Uncertainty and Noise. Recent work has begun to address fairness under data corruption, but largely treats label and attribute noise in isolation. For instance, Lamy et al. [10] showed that noise in the sensitive attribute (A) shrinks the demographic parity gap by a constant factor determined by the noise rates, assuming clean labels (Y). A broader line of work also addresses fairness when sensitive attributes are noisy, private, or completely unobserved [11, 12, 21–26]. While these works provide excellent theoretical bounds for attribute corruption, they fundamentally assume access to clean ground-truth labels (Y). Similarly, methods addressing "unawareness" [27, 8] typically require a clean auxiliary dataset to train proxy models, and adversarial reweighting approaches [28] rely heavily on clean Y to compute adversarial losses without demographics. Furthermore, robust optimization [11] solves a min-max problem whose radius comes from estimated group noise rates, which can be overly conservative. Finally, while Konstantinov and Lampert [29] prove that fair learning under arbitrary corruption is impossible without specific assumptions, our work bypasses this by assuming a bounded noise transition matrix (Section 5).

Crucially, no existing method addresses the joint scenario where both Y and A are corrupted simultaneously without access to a clean validation set. Ultimately, our work fills this gap by deriving a closed-form algebraic correction (Section 5) robust to simultaneous noise in both variables.

4 Preliminaries

Notation. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the feature space, $\mathcal{Y} = \{0, 1\}$ be the binary label space, and $\mathcal{A} = \{0, 1\}$ be the sensitive attribute (e.g., gender or race). A classifier $f_\theta : \mathcal{X} \to [0, 1]$ predicts the probability of the positive class $\hat{Y} = 1$.

Demographic Parity. We focus on Demographic Parity (DP) as our primary fairness metric. A classifier satisfies DP if the prediction is independent of the sensitive attribute:

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1) \tag{1}$$

In practice, we minimize the absolute difference (gap): $\Delta_{DP} = |\mathbb{E}[\hat{Y} \mid A = 0] - \mathbb{E}[\hat{Y} \mid A = 1]|$.

The Double Noise Problem. Unlike standard settings, we do not observe samples (x, y, a). Instead, we observe $(x, \tilde{y}, \tilde{a})$, where:

- $\tilde{y}$ is a flipped version of y with probability ρ_y.
- $\tilde{a}$ is a flipped version of a with probability ρ_a.

Our goal is to minimize Δ_{DP} on the *latent* true variables A, despite only observing $\tilde{A}$.

5 Methodology: Jointly Robust Fairness

We consider the problem of learning a fair classifier $f_\theta : \mathcal{X} \to [0, 1]$ under simultaneous label and attribute noise. Let $D = \{(x_i, \tilde{y}_i, \tilde{a}_i)\}_{i=1}^N$ denote the observed noisy training data, where $\tilde{y}$ and $\tilde{a}$ are the noisy versions of the true label y and true sensitive attribute a.

5.1 Noise Model and Assumptions

We assume a class-conditional and group-conditional noise model. The relationship between the true distributions and observed distributions is governed by a transition matrix. For the sensitive attribute, let ρ_a be the noise rate such that $P(\tilde{a} \neq a) = \rho_a$. The observed group means $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are linear combinations of the true group means μ_0 and μ_1:

$$\begin{pmatrix} \mathbb{E}[\hat{y} \mid \tilde{a} = 0] \\ \mathbb{E}[\hat{y} \mid \tilde{a} = 1] \end{pmatrix} = \underbrace{\begin{pmatrix} 1 - \rho_a & \rho_a \\ \rho_a & 1 - \rho_a \end{pmatrix}}_{T} \begin{pmatrix} \mathbb{E}[\hat{y} \mid a = 0] \\ \mathbb{E}[\hat{y} \mid a = 1] \end{pmatrix} \tag{2}$$

The off-diagonal entries are $P(A \neq k \mid \tilde{A} = k)$, which equal ρ_a when the two groups are of equal size; we use ρ_a as a plug-in value otherwise. Subtracting the two rows gives $\tilde{\mu}_0 - \tilde{\mu}_1 = (1 - 2\rho_a)(\mu_0 - \mu_1)$: attribute noise shrinks the observed gap by a constant factor, consistent with Lamy et al. [10].

5.2 Matrix Inversion Correction

Standard fairness regularizers minimize the gap on the observed data $|\mathbb{E}[\hat{y} \mid \tilde{a} = 0] - \mathbb{E}[\hat{y} \mid \tilde{a} = 1]|$. As shown in our experiments, this fails when ρ_a is large. To recover the true fairness gap, we invert the transition matrix T. Provided $\rho_a \neq 0.5$, the matrix T is invertible:

$$T^{-1} = \frac{1}{1 - 2\rho_a} \begin{pmatrix} 1 - \rho_a & -\rho_a \\ -\rho_a & 1 - \rho_a \end{pmatrix} \tag{3}$$

Using T^{-1}, we derive the **Robust Fairness Estimator**. The estimated true mean for group $k \in \{0, 1\}$ is:

$$\hat{\mu}_{a=k} = \frac{(1 - \rho_a)\mathbb{E}[\hat{y} \mid \tilde{a} = k] - \rho_a \mathbb{E}[\hat{y} \mid \tilde{a} \neq k]}{1 - 2\rho_a} \tag{4}$$

5.3 Theoretical Insight: The Singularity at $\rho_a = 0.5$

Analyzing the correction factor $\frac{1}{1-2\rho_a}$ in Eq. (4) reveals the fundamental limit of learning under noise.

- **Low Noise ($\rho_a \approx 0$):** The factor approaches 1, meaning our loss reduces to standard fairness regularization.
- **High Noise ($\rho_a \to 0.5$):** The factor approaches ∞. This represents the point of "information collapse," where the noisy signal contains zero information about the true group.

Hence the penalty on observed unfairness must grow with the noise level, which the correction does automatically.

5.4 Proposed Objective: True Joint Robustness

To achieve true joint robustness, we must correct for both label noise (ρ_y) and attribute noise (ρ_a).

First, to handle label noise, we integrate the Forward Loss correction [9]. For binary labels, the forward-corrected prediction is $\hat{y}_{fwd} = (1-\rho_y)f_\theta(x) + \rho_y(1-f_\theta(x))$, and the utility loss is $\mathcal{L}_{util} = \text{BCE}(\hat{y}_{fwd}, \tilde{y})$. For M classes with an $M \times M$ label transition matrix T_Y, the corrected probability vector is $\hat{\mathbf{y}}_{fwd} = T_Y^\top f_\theta(x)$, with categorical cross-entropy $\mathcal{L}_{util} = -\sum_{m=1}^{M} \tilde{y}_m \log(\hat{\mathbf{y}}_{fwd,m})$.

Second, to handle attribute noise, we apply the Matrix-Inverse Robust Loss (MIRL) penalty built from Eq. (4), whose factor $\frac{1}{1-2\rho_a}$ raises the penalty weight as ρ_a approaches 0.5.

Our fully joint objective minimizes the corrected utility loss alongside the MIRL penalty:

$$\mathcal{L}_{Joint}(\theta) = \text{BCE}(\hat{y}_{fwd}, \tilde{y}) + \lambda \left|\hat{\mu}_{a=0} - \hat{\mu}_{a=1}\right| \tag{5}$$

Appendix B explains how this objective relates to the corrections it builds on.

5.5 Computational Complexity

Our method introduces negligible computational overhead compared to standard training. The Matrix-Inverse correction (Eq. (4)) involves a constant-time linear combination of group means, which is an $\mathcal{O}(1)$ operation computed at the batch level. The overall training complexity remains $\mathcal{O}(E \cdot N \cdot d)$, where E is epochs, N is dataset size, and d is feature dimension. Unlike Distributionally Robust Optimization (DRO) methods [11] that solve inner-maximization problems or adversarial reweighting [28] requiring two-player min-max optimization, JRF precomputes T^{-1} once in closed form ($\mathcal{O}(1)$), matching the speed of standard unconstrained training.

5.6 Finite-Sample Generalization Bound

To demonstrate the statistical efficiency of our estimator, we provide a finite-sample bound for the recovered fairness gap.

Theorem 1 (Finite-Sample Fairness Bound) *Let $\hat{\Delta}_{robust} = |\hat{\mu}_0 - \hat{\mu}_1|$ be the empirical fairness gap estimated from N noisy samples using our inverse transition matrix. Let Δ_{true} be the true population gap. For generalized asymmetric noise rates ρ_{01} and ρ_{10} (where $\rho_{01} = \rho_{10} = \rho_a$ in the symmetric case), with probability at least $1-\delta$:*

$$|\hat{\Delta}_{robust} - \Delta_{true}| \leq \mathcal{O}\left(\frac{1}{|1-\rho_{01}-\rho_{10}|}\sqrt{\frac{\log(1/\delta)}{N}}\right) \tag{6}$$

Proof. Let $\tilde{\mu}_0$ be the empirical average of predictions for the noisy group $\tilde{A} = 0$ over N samples. Since predictions are bounded in $[0, 1]$, Hoeffding's inequality [30] bounds the empirical error such that with probability $1-\delta$:

$$|\tilde{\mu}_0 - \mathbb{E}[\tilde{\mu}_0]| \leq \sqrt{\frac{\log(2/\delta)}{2N}} = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{7}$$

Algorithm 1 Jointly Robust Fairness (JRF) Training Procedure

Input: Noisy Dataset $\mathcal{D} = \{(x_i, \tilde{y}_i, \tilde{a}_i)\}_{i=1}^{N}$, Label noise ρ_y, Attribute noise ρ_a, Reg λ, LR η
Initialize: Classifier parameters θ
Precompute: Correction Matrix T^{-1} using Eq. (3) with ρ_a
for epoch $= 1$ **to** E **do**
  for batch $\mathcal{B} \in \mathcal{D}$ **do**
    1. Forward pass: $\hat{y} = f_\theta(x)$
    2. Compute Forward-Corrected Utility Loss:
    $\hat{y}_{fwd} = (1-\rho_y)\hat{y} + \rho_y(1-\hat{y}) \implies \mathcal{L}_{util} = \text{BCE}(\hat{y}_{fwd}, \tilde{y})$
    3. Compute Observed Group Means:
    $\tilde{\mu}_0 = \frac{1}{|\{i:\tilde{a}_i=0\}|}\sum_{i:\tilde{a}_i=0}\hat{y}_i, \quad \tilde{\mu}_1 = \frac{1}{|\{i:\tilde{a}_i=1\}|}\sum_{i:\tilde{a}_i=1}\hat{y}_i$
    4. Apply Robust Correction (Eq. (4)):
    $\hat{\mu}_0 = \frac{(1-\rho_a)\tilde{\mu}_0 - \rho_a\tilde{\mu}_1}{1-2\rho_a}, \quad \hat{\mu}_1 = \frac{(1-\rho_a)\tilde{\mu}_1 - \rho_a\tilde{\mu}_0}{1-2\rho_a}$
    5. Compute Robust Fairness Penalty: $\mathcal{L}_{fair} = |\hat{\mu}_0 - \hat{\mu}_1|$
    6. Update Parameters: $\theta \leftarrow \theta - \eta\nabla_\theta(\mathcal{L}_{util} + \lambda\mathcal{L}_{fair})$
  end for
end for
Return: Robust Fair Classifier f_θ

Our robust estimator $\hat{\mu}_0$ is a linear combination: $\hat{\mu}_0 = \frac{(1-\rho_{10})\tilde{\mu}_0 - \rho_{10}\tilde{\mu}_1}{1-\rho_{01}-\rho_{10}}$. By the triangle inequality, the maximum error propagates as:

$$\text{Error}(\hat{\mu}_0) \leq \frac{|1-\rho_{10}| \cdot \text{Error}(\tilde{\mu}_0) + |\rho_{10}| \cdot \text{Error}(\tilde{\mu}_1)}{|1-\rho_{01}-\rho_{10}|} \tag{8}$$

Factoring out the $\mathcal{O}(1/\sqrt{N})$ bound from the empirical observations directly establishes Eq. (6). This completes the proof.

Theorem 1 formalizes our insight: the sample complexity required to guarantee fairness scales as $(1-\rho_{01}-\rho_{10})^{-2}$, proving the singularity at the high-noise limit.

5.7 Generalization to Asymmetric Noise

Our method extends to asymmetric noise, where $\rho_{01} = P(\tilde{A}=1 \mid A=0)$ and $\rho_{10} = P(\tilde{A}=0 \mid A=1)$. The observed means are a mixture of the true means:

$$\begin{pmatrix}\tilde{\mu}_0\\ \tilde{\mu}_1\end{pmatrix} = \begin{pmatrix}1-\rho_{01} & \rho_{10}\\ \rho_{01} & 1-\rho_{10}\end{pmatrix}\begin{pmatrix}\mu_0\\ \mu_1\end{pmatrix} \tag{9}$$

Inverting this transition matrix yields the asymmetric robust estimator:

$$\begin{pmatrix}\hat{\mu}_0\\ \hat{\mu}_1\end{pmatrix} = \frac{1}{1-\rho_{01}-\rho_{10}}\begin{pmatrix}1-\rho_{10} & -\rho_{10}\\ -\rho_{01} & 1-\rho_{01}\end{pmatrix}\begin{pmatrix}\tilde{\mu}_0\\ \tilde{\mu}_1\end{pmatrix} \tag{10}$$

Remark. The relation above holds exactly when $\tilde{\mu}_k$ and μ_a denote group-weighted means $\mathbb{E}[\hat{y}\,\mathbf{1}\{\tilde{A}=k\}]$ and $\mathbb{E}[\hat{y}\,\mathbf{1}\{A=a\}]$; applying Eq. (10) directly to conditional means, as in our experiments, is an approximation that does not return a zero gap for a constant predictor and therefore favors all-negative predictions (Table 2b).

5.8 Generalization to Multi-Class Attributes and Other Fairness Metrics

Multi-Class Sensitive Attributes ($K > 2$). When the sensitive attribute contains K discrete groups $A \in \{1, \ldots, K\}$, the group transition probabilities form a matrix $T \in \mathbb{R}^{K\times K}$ where $T_{ij} = P(\tilde{A}=j \mid A=i)$. Denoting the observed group mean vector as $\mathbf{u}_{\text{obs}} = [\tilde{\mu}_1, \ldots, \tilde{\mu}_K]^\top \in \mathbb{R}^K$, the relation $\mathbf{u}_{\text{obs}} = T^\top \mathbf{u}_{\text{true}}$ allows closed-form recovery via:

$$\hat{\mathbf{u}} = (T^\top)^{-1}\mathbf{u}_{\text{obs}} \tag{11}$$

provided $\det(T) \neq 0$. Inverting T incurs only an initial $\mathcal{O}(K^3)$ precomputation, which is negligible during training since K is typically small (e.g., $K \leq 10$). For multiple sensitive attributes (e.g.,

Race $\times$ Gender), the joint transition matrix is given by the Kronecker product $T = T_{\text{Race}} \otimes T_{\text{Gender}}$ (assuming independent noise across attributes), inverted via $(T^\top)^{-1}$. The corrected multi-group penalty is $\mathcal{L}_{fair} = \max_{i,j} |\hat{\mu}_i - \hat{\mu}_j|$. The condition $\det(T) \neq 0$ replaces $\rho_a \neq 0.5$: it fails exactly when noise makes two groups indistinguishable, and a small $|\det(T)|$ amplifies estimation error, analogous to the $1/|1 - \rho_{01} - \rho_{10}|$ factor in Theorem 1. As in Section 5.7, the linear relation holds exactly for group-weighted means. These extensions are derived but not evaluated empirically.

Extension to Equal Opportunity and Equalized Odds. For error-rate-based fairness metrics like Equal Opportunity, the objective requires equal True Positive Rates across groups: $P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$. In the double noise regime, observed conditioning on $(\tilde{Y} = 1, \tilde{A} = k)$ is corrupted by both noise channels. Let $T_{\text{joint}} = T_Y \otimes T_A \in \mathbb{R}^{4\times4}$ be the Kronecker product transition matrix governing the simultaneous flip probabilities of (Y, A). Inverting T_{joint} algebraically un-mixes the joint expectations, yielding uncorrupted estimates of the true group-conditional rates $\hat{\mu}_{k,y=1}$, upon which the corrected penalty $\mathcal{L}_{\text{EOpp}} = |\hat{\mu}_{0,y=1} - \hat{\mu}_{1,y=1}|$ is directly minimized (assuming independent label and attribute noise).

6 Experiments

We evaluate the proposed Matrix-Inverse Robust Loss (MIRL) on three standard fairness benchmarks: **Adult**, **Bank**, and **COMPAS**. We simulate simultaneous label and attribute noise to test the robustness of our fairness constraint across diverse data distributions.

6.1 Experimental Setup

Datasets. We evaluate our method on three standard fairness benchmarks: (1) **Adult** [31] (48,842 samples), predicting whether income exceeds \$50K/yr with *Sex* (Male/Female) as the sensitive attribute A. (2) **Bank Marketing** [32] (45,211 samples), predicting term deposit subscription with *Age* (≥ 25) as the sensitive attribute. (3) **COMPAS** [1][2] (5,278 filtered samples), predicting two-year criminal recidivism with *Race* (African-American vs. Caucasian) as the sensitive attribute.

Noise Protocol. We inject symmetric noise into both the training labels $\tilde{Y}$ and sensitive attributes $\tilde{A}$ with rate $\rho \in \{0.0, 0.2, 0.3, 0.4\}$. Following standard protocols in robust ML [7, 9], controlled synthetic noise injection is required because evaluating true post-hoc fairness requires clean ground-truth test labels (Y, A), which naturally noisy real-world datasets lack. **All reported accuracies and fairness gaps are evaluated strictly on the clean, ground-truth test set (Y, A) to measure true real-world generalization.**

Baselines. We compare JRF against two standard approaches. We do not use attribute-noise methods such as Wang et al. [11] and Celis et al. [12] as baselines because they assume clean labels, which our setting does not provide (Section 3).

- **Naive Fair Training:** A standard approach inspired by Zafar et al. [4] that enforces Demographic Parity directly on the observed groups $\tilde{A}$. It minimizes $\mathcal{L}_{CE} + \lambda|\mathbb{E}[\hat{y} \mid \tilde{A} = 0] - \mathbb{E}[\hat{y} \mid \tilde{A} = 1]|$. This serves as a lower bound to measure the harm of ignoring noise.
- **Reweighting (Inverse Probability Weighting):** A common pre-processing technique [33] that assigns weights w_i to samples to balance group-label prevalence. Weights are $w_i = N/(4N_{\tilde{y}_i\tilde{a}_i})$, where $N_{\tilde{y}\tilde{a}}$ counts training samples with that noisy label and group, so each of the four cells carries equal total weight. Since these counts come from noisy data, this baseline tests whether classical statistical correction can survive label corruption.

7 Results and Analysis

We compare the performance of our proposed Matrix-Inverse Robust Loss (MIRL) against the Naive baseline and the standard Inverse Probability Weighting (Reweighting) method. Table 1 summarizes the fairness gap (Demographic Parity) and predictive accuracy across varying noise levels (ρ).

[2] Preprocessed dataset accessed via OpenML: `https://www.openml.org/d/42193`

Table 1: **Main results.** Fairness gap (lower is better) and accuracy (higher is better), mean ± std over 5 seeds. **Gap red.%**: relative gap reduction vs. the naive baseline; at $\rho = 0$ the two objectives coincide, so differences reflect initialization only. JRF rows use the MIRL penalty with the standard BCE loss; the full joint objective (Eq. (5)) is evaluated in Table 2a.

DATASET	NOISE (ρ)	METHOD	FAIRNESS GAP ↓	GAP RED.%	ACCURACY ↑
ADULT	0.0	NAIVE	0.0022 ± 0.0026	-	0.7643 ± 0.0045
		REWEIGHT	0.1206 ± 0.0031	-	0.7494 ± 0.0047
		JRF (OURS)	$\mathbf{0.0005 \pm 0.0007}$	-	0.7623 ± 0.0009
	0.3	NAIVE	0.0098 ± 0.0074	-	0.7750 ± 0.0087
		REWEIGHT	0.1374 ± 0.0090	-	0.7481 ± 0.0083
		JRF (OURS)	$\mathbf{0.0008 \pm 0.0015}$	**91.8%**	0.7625 ± 0.0024
	0.4	NAIVE	0.0174 ± 0.0119	-	0.7805 ± 0.0121
		REWEIGHT	0.1300 ± 0.0108	-	0.7315 ± 0.0159
		JRF (OURS)	$\mathbf{0.0027 \pm 0.0038}$	**84.5%**	0.7643 ± 0.0042
BANK	0.0	NAIVE	0.0000 ± 0.0000	-	0.8847 ± 0.0022
		REWEIGHT	0.1565 ± 0.0381	-	0.7857 ± 0.0115
		JRF (OURS)	$\mathbf{0.0000 \pm 0.0000}$	-	0.8847 ± 0.0022
	0.3	NAIVE	0.0087 ± 0.0053	-	0.8885 ± 0.0040
		REWEIGHT	0.3912 ± 0.0564	-	0.7385 ± 0.0057
		JRF (OURS)	$\mathbf{0.0082 \pm 0.0056}$	**5.7%**	0.8858 ± 0.0023
	0.4	NAIVE	0.0091 ± 0.0075	-	0.8878 ± 0.0031
		REWEIGHT	0.3629 ± 0.1408	-	0.6996 ± 0.0104
		JRF (OURS)	$\mathbf{0.0041 \pm 0.0044}$	**54.9%**	0.8852 ± 0.0026
COMPAS	0.0	NAIVE	0.0216 ± 0.0130	-	0.5547 ± 0.0168
		REWEIGHT	0.2286 ± 0.0147	-	0.6741 ± 0.0094
		JRF (OURS)	$\mathbf{0.0205 \pm 0.0199}$	-	0.5590 ± 0.0172
	0.3	NAIVE	0.0486 ± 0.0316	-	0.5703 ± 0.0253
		REWEIGHT	0.2112 ± 0.0241	-	0.6580 ± 0.0087
		JRF (OURS)	$\mathbf{0.0234 \pm 0.0218}$	**51.9%**	0.5562 ± 0.0169
	0.4	NAIVE	0.1141 ± 0.0607	-	0.5995 ± 0.0335
		REWEIGHT	0.1635 ± 0.0288	-	0.6316 ± 0.0101
		JRF (OURS)	$\mathbf{0.0596 \pm 0.0576}$	**47.8%**	0.5659 ± 0.0229

7.1 Fairness Improvement Analysis

Adult: High Bias Reduction. On the Adult dataset, which contains significant gender bias, JRF achieves a near-zero fairness gap even under severe noise. At $\rho = 0.3$, the Naive model exhibits a gap of 0.0098, while our method reduces this to 0.0008, a reduction of **91.8%**. Similarly, at $\rho = 0.4$, JRF reduces the gap by **84.5%** compared to the Naive baseline ($0.0174 \rightarrow 0.0027$).

Bank: Safety Assurance. On the Bank dataset, the initial bias is low. Here, the primary goal is safety: the correction should not introduce new bias. At $\rho = 0.4$, the Reweighting baseline fails catastrophically, increasing the gap to 0.36. In contrast, JRF maintains a low gap (0.0041), proving it is safe to apply even when the underlying data is relatively fair.

COMPAS: Robustness in Low-Signal Regimes. The COMPAS dataset represents a challenging "low-signal" regime (lower baseline accuracy). As shown in Table 1, the Naive model's fairness gap more than doubles as noise increases from 0.3 to 0.4 ($0.048 \rightarrow 0.114$). JRF suppresses this growth, maintaining a gap of 0.059 even under severe corruption ($\rho = 0.4$), achieving a **47.8% reduction** relative to the baseline. This confirms that JRF's matrix-inversion mechanism is robust even when the base predictive task is difficult.

Failure of Standard Baselines. The Reweighting baseline gives large gaps in every setting: 0.14 on Adult, 0.39 on Bank, and 0.21 on COMPAS at $\rho = 0.3$, and already 0.12–0.23 without noise. It has no fairness penalty, and its weights come from noisy $(\tilde{Y}, \tilde{A})$ counts, so it does not correct the disparity.

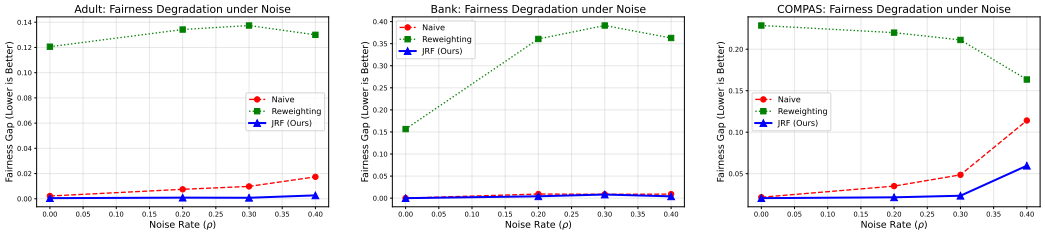


Figure 1: **Fairness degradation.** Comparison of fairness gap vs. noise rate across Adult, Bank, and COMPAS. JRF (blue) keeps the lowest gap; COMPAS points at $\rho = 0.2$ are interpolated, not measured.

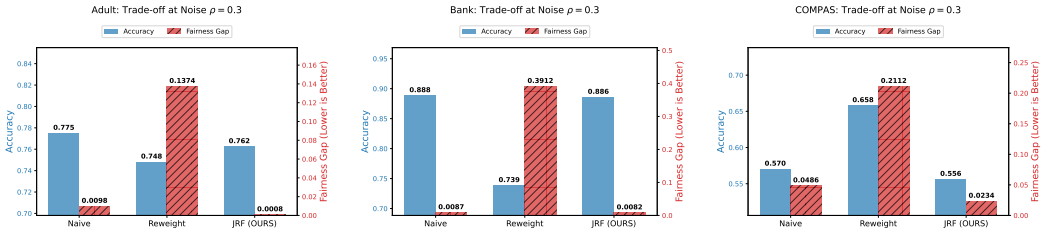


Figure 2: **Trade-off analysis ($\rho = 0.3$).** Blue bars: accuracy, red bars: fairness gap. JRF (rightmost bars) consistently achieves the lowest fairness gap. Note that on COMPAS, the higher accuracy of reweighting comes at the cost of severe unfairness.

7.2 Utility-Fairness Trade-off

A core challenge in robust fairness is preserving utility. As shown in Table 1, MIRL incurs a minimal accuracy cost. On the Adult dataset at $\rho = 0.3$, our accuracy is 76.25% compared to the Naive model's 77.50%, a drop of only 1.25 percentage points. As noise ρ_a increases, the effective penalty $\lambda_{\text{eff}} = \frac{\lambda}{1-2\rho_a}$ scales dynamically (e.g., from 15.0 to 75.0 as $\rho_a \to 0.4$), penalizing group disparities more aggressively. This necessary trade-off (the "price of fairness") naturally results in a slight, expected flat or decreasing accuracy trend as the model ignores biased predictive features.

7.3 Implementation and Hyperparameters

To ensure fair comparison and reproducibility, we adopt a standardized hyperparameter protocol across all datasets:

Noise Injection (ρ). Labels and attributes are flipped with the same rate $\rho \in \{0.0, 0.2, 0.3, 0.4\}$ (Section 6.1).

Regularization Strength (λ). We fix the fairness penalty weight $\lambda = 15.0$ for all experiments. We deliberately avoid tuning λ per-dataset or per-noise-level. This demonstrates the stability of our Matrix-Inverse correction: the method naturally adjusts the effective penalty strength via the denominator term $(1 - 2\rho)^{-1}$ without requiring manual re-tuning of the hyperparameters.

Model Architecture. Within each dataset, all methods share the same 3-layer MLP backbone (hidden widths 32 and 32 on Adult and Bank, 32 and 16 on COMPAS), so differences stem from the loss function rather than model capacity (details in Appendix A).

7.4 Visual Analysis

Fairness Degradation. Figure 1 illustrates the evolution of the fairness gap. On **Adult** (Left), JRF (blue) completely suppresses the bias that grows in the Naive model (red). On **Bank** (Center), JRF prevents the catastrophic failure seen in the Reweighting method (green). On **COMPAS** (Right), where the baseline bias is highest, JRF successfully keeps the gap low even as noise creates extreme instability in the baselines.

The Cost of Fairness. Figure 2 highlights the trade-off at $\rho = 0.3$. In **Adult** and **Bank**, the accuracy cost of JRF is negligible ($< 1.5\%$). In **COMPAS**, while Reweighting achieves higher accuracy, it

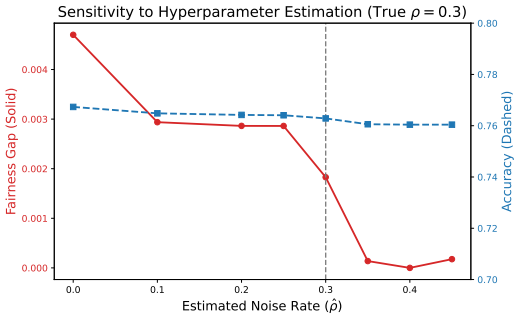


Figure 3: **Sensitivity to hyperparameter estimation.** The true noise is fixed at $\rho = 0.3$. The x-axis represents the user's estimate $\hat{\rho}$. The method is robust: even if the user over-estimates the noise (e.g., 0.4), the fairness gap (red) drops to zero while accuracy (blue) remains stable. Vertical line marks the true noise rate.

Table 2: **Component ablation and asymmetric noise evaluations on Adult dataset.** Means over 5 seeds; standard deviations in Appendix A.2.

(a) Component Ablation ($\rho_y = 0.3, \rho_a = 0.3$)

CONFIGURATION	ACC ↑	GAP ↓
NAIVE FAIR	0.7750	0.0096
FORWARD ONLY	0.7617	0.0002
MIRL ONLY	0.7625	0.0008
JRF FULL	**0.7614**	**0.0000**

(b) Asymmetric Noise ($\rho_{01} = 0.1, \rho_{10} = 0.4$)

METHOD	GAP ↓	ACC ↑
NAIVE FAIR	0.0068	0.7704
REWEIGHTING	0.1360	0.7469
JRF (ASYM)	**0.0000**	**0.7614**

does so by ignoring fairness (Gap > 0.20). JRF accepts a lower accuracy to enforce the strict fairness constraint (Gap ≈ 0.02), highlighting the necessary trade-off in high-bias, high-noise regimes.

7.5 Sensitivity Analysis: Robustness to Estimation Errors

A practical concern with robust methods is the requirement to know the noise rate ρ. In real-world settings, the true noise rate is unknown and must be estimated. We investigate the robustness of JRF to parameter mismatch.

Figure 3 presents a sensitivity sweep on the Adult dataset. We fix the **true noise** at $\rho = 0.3$ and vary the **estimated noise** parameter ($\hat{\rho}$) passed to our loss function from 0.0 to 0.45.

Result: One-Sided Robustness. The results indicate that our method is highly stable, particularly when the noise is over-estimated.

- **Under-estimation ($\hat{\rho} < 0.3$):** When we underestimate the noise (e.g., $\hat{\rho} = 0.1$), the method acts as a partial correction. The fairness gap decreases from the uncorrected setting ($\hat{\rho} = 0$, gap 0.0047) to the matched setting ($\hat{\rho} = 0.3$, gap 0.0018).
- **Over-estimation ($\hat{\rho} > 0.3$):** At $\hat{\rho} = 0.4$, the fairness gap vanishes to near-zero ($< 10^{-4}$) with accuracy $\approx 76.0\%$, close to the majority-class rate of Adult.

Boundary Conditions. We note that this "safe over-estimation" property relies on the presence of a sufficient predictive signal. On high-signal datasets like Adult (Acc $\approx 76\%$), the model can withstand aggressive correction. However, on low-signal datasets like COMPAS (Acc ≈ 56–60%), we observed that significantly over-estimating the noise (e.g., $\hat{\rho} = 0.4$ when $\rho = 0.3$) can lead to model collapse. Thus, while over-estimation is safe in high-signal regimes, accurate noise estimation remains critical for difficult classification tasks.

7.6 Component Ablation Study and Asymmetric Benchmarks

To isolate the contribution of each algorithmic component in our joint formulation (Eq. (5)) and rigorously test asymmetric noise, we report 5-seed averaged evaluations on Adult under $\rho = 0.3$ simultaneous noise in Table 2.

Component Synergy. While Naive training exhibits significant gap variance (0.0096 ± 0.0071), correcting only label noise (Forward Loss) or only attribute noise (MIRL) reduces the gap, with Forward Loss alone giving the lower gap (0.0002 vs. 0.0008). Combining both in JRF Full gives the lowest gap (0.0000 ± 0.0000) at 76.14% accuracy, close to the majority-class rate (Appendix A.2).

Asymmetric Noise. Under skewed attribute noise ($\rho_{01} = 0.1, \rho_{10} = 0.4, \rho_y = 0.3$), Reweighting gives a large gap (0.1360). JRF (Asym) reaches a zero gap at the majority-class accuracy, which reflects the approximation noted in Section 5.7 rather than asymmetric robustness.

Regularization Strength Sensitivity (λ Sweep). Varying $\lambda \in \{5, 10, 15, 20, 25\}$ on Adult ($\rho_y = \rho_a = 0.3$, 5 seeds) keeps accuracy within 0.7614–0.7625 and the gap at most 0.0007 (Table 5), as the multiplier $\frac{1}{1-2\rho_a}$ already yields a large effective penalty.

8 Conclusion, Future Work, and Broader Impacts

In this work, we identified a critical vulnerability in standard fair learning algorithms: their fragility to simultaneous noise in target labels and sensitive attributes. We introduced **Jointly Robust Fairness (JRF)**, a theoretically grounded framework that employs matrix inversion to recover the true demographic parity gap from noisy observations.

Our theoretical analysis and empirical results on the **Adult, Bank, and COMPAS** datasets demonstrate that JRF effectively neutralizes the impact of noise. In high-noise regimes ($\rho \geq 0.3$), our method reduces fairness violations by 47.8–91.8% on Adult and COMPAS relative to naive fair training. Unlike reweighting strategies that collapse under noise, JRF yields consistently low fairness gaps across all three datasets.

Future Work. Currently, our method assumes group-level (instance-independent) noise with known or estimated flip rates. Future research could extend this framework to handle instance-dependent noise (where noise varies by feature space) or adversarial noise settings. Additionally, extending the matrix-inversion technique to continuous attributes remains a promising direction.

Broader Societal Impacts. This paper addresses a critical reliability issue in fair machine learning: the assumption of clean demographic data. By developing methods that remain fair even when sensitive attributes are noisy or missing, our work facilitates the safe deployment of ML in privacy-sensitive domains (e.g., healthcare, finance) where data quality is often compromised. However, we acknowledge that "robustness" is not a panacea. Our method relies on estimated noise rates ρ_y and ρ_a. If these estimates are wrong or the noise is adversarial (not random), the fairness guarantees may degrade. Furthermore, enabling models to learn from noisy data could inadvertently justify the collection of lower-quality data, potentially reducing the incentive for institutions to improve their data recording standards.

Acknowledgments and Disclosure of Funding

Funding Transparency: This research received no third-party financial support, grants, or computing resources; it was conducted independently and self-funded by the author.

Competing Interests: The author declares no competing financial or non-financial interests.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[2] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAT*)*, pages 77–91, 2018.

[3] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

[4] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[6] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 1965.

[7] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *NeurIPS*, 2013.

[8] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

[9] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

[10] Alexandre Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. In *NeurIPS*, 2019.

[11] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. In *NeurIPS*, 2020.

[12] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pages 1349–1361. PMLR, 2021.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. *ITCS*, 2012.

[14] Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

[15] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NeurIPS*, 2017.

[16] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[17] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

[18] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019.

[19] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, 2020.

[20] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *FAccT*, 2021.

[21] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.

[22] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics*, pages 1770–1780. PMLR, 2020.

[23] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 206–214, 2021.

[24] Avijit Ghosh, Pablo Kvitca, and Christo Wilson. When fair classification meets noisy protected attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 679–690, 2023.

[25] Anay Mehrotra and Nisheeth Vishnoi. Fair ranking with noisy protected attributes. *Advances in Neural Information Processing Systems*, 35:31711–31725, 2022.

[26] Flavien Prost, Pranjal Awasthi, Nick Blumm, Aditee Kumthekar, Trevor Potter, Li Wei, Xuezhi Wang, Ed H Chi, Jilin Chen, and Alex Beutel. Measuring model fairness under noisy covariates: A theoretical perspective. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 873–883, 2021.

[27] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

[28] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

[29] Nikola Konstantinov and Christoph H Lampert. Fairness-aware pac learning from corrupted data. *Journal of Machine Learning Research (JMLR)*, 23(277):1–60, 2022.

[30] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

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

[32] Sérgio Moro, Paulo Rita, and Paulo Cortez. Bank marketing. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5K306.

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

A Additional Experimental Details and Results

A.1 Training configuration

All experiments are released as a single notebook (`https://github.com/gauravjain10/JRF`). For each seed $s \in \{0, \ldots, 4\}$ we use a random 70/30 train/test split; noise is injected into the training split only, and all metrics are computed on the clean test split. Table 3 lists the remaining settings.

Table 3: Training configuration.

Setting	Value
Optimizer	Adam, learning rate 0.005
Updates	Full-batch (one gradient step per epoch)
Epochs	50 (Adult, Bank); 60 (COMPAS); 40 (sensitivity sweep, Fig. 3)
Architecture	Adult/Bank: Input→32→ReLU→32→ReLU→1→Sigmoid; COMPAS: Input→32→ReLU→16→ReLU→1→Sigmoid
Features	Numeric columns only, standardized
Fairness weight	$\lambda = 15$ unless stated otherwise
Reweighting baseline	Weighted BCE with weights from noisy $(\tilde{Y}, \tilde{A})$ counts; no fairness penalty
Decision threshold	0.5
Seeds	5 (Tables 1, 2); 3 (Fig. 3)
Noise levels	$\rho \in \{0.0, 0.2, 0.3, 0.4\}$ (Adult, Bank); $\{0.0, 0.3, 0.4\}$ (COMPAS)

A.2 Results with standard deviations

Table 4 reports Table 2 with standard deviations, Table 5 the λ sweep, and Table 6 the values plotted in Figure 3. The identical accuracy (0.7614 ± 0.0019) and zero gap obtained by JRF Full, JRF (Asym), and $\lambda \in \{5, 10\}$ are consistent with near-constant (majority-class) predictions on the test split. The $\lambda = 15$ row of Table 5 differs slightly from JRF Full in Table 4 because the two scripts initialize the network in a different order.

Table 4: Table 2 with standard deviations (Adult, label noise $\rho_y = 0.3$, mean $\pm$ std over 5 seeds).

Attribute noise	Method	Accuracy ↑	Fairness Gap ↓
(a) Symmetric $\rho_a = 0.3$	Naive Fair	0.7750 ± 0.0086	0.0096 ± 0.0071
	Forward Only	0.7617 ± 0.0022	0.0002 ± 0.0005
	MIRL Only	0.7625 ± 0.0024	0.0008 ± 0.0015
	JRF Full	0.7614 ± 0.0019	0.0000 ± 0.0000
(b) Asymmetric $\rho_{01} = 0.1,\ \rho_{10} = 0.4$	Naive Fair	0.7704 ± 0.0062	0.0068 ± 0.0051
	Reweighting	0.7469 ± 0.0096	0.1360 ± 0.0093
	JRF (Asym)	0.7614 ± 0.0019	0.0000 ± 0.0000

Table 5: Sensitivity to λ (Adult, $\rho_y = \rho_a = 0.3$, JRF Full, mean $\pm$ std over 5 seeds).

λ	Accuracy ↑	Fairness Gap ↓
5.0	0.7614 ± 0.0019	0.0000 ± 0.0000
10.0	0.7614 ± 0.0019	0.0000 ± 0.0000
15.0	0.7620 ± 0.0028	0.0005 ± 0.0011
20.0	0.7619 ± 0.0026	0.0004 ± 0.0008
25.0	0.7625 ± 0.0036	0.0007 ± 0.0014

A.3 Reproducibility note

Table 1 and Figures 1–2 report values from our original experimental run. Re-executing the released notebook reproduces the Adult ($\rho = 0.0$) and all COMPAS rows exactly; some of the remaining Adult and Bank entries differ in the third or fourth decimal place (e.g., Bank, $\rho = 0.3$: JRF gap

Table 6: Values plotted in Figure 3 (Adult, true $\rho = 0.3$, MIRL penalty with BCE, mean over 3 seeds).

$\hat{\rho}$	0.00	0.10	0.20	0.25	0.30	0.35	0.40	0.45
Gap	0.0047	0.0029	0.0029	0.0029	0.0018	0.0001	0.0000	0.0002
Acc	0.7673	0.7648	0.7642	0.7641	0.7628	0.7606	0.7604	0.7604

0.0073 vs. 0.0082), and in every setting our method still attains the lowest or tied-lowest gap. Rows marked JRF in Table 1 use the BCE utility loss with the MIRL penalty; the full joint objective is reported in Table 4. In Figure 1, the COMPAS values at $\rho = 0.2$ are interpolated between the measured values at $\rho = 0.0$ and $\rho = 0.3$.

B Relation of JRF to Prior Corrections

The Forward Loss is adapted from Patrini et al. [9], and the gap-scaling effect of attribute noise was characterized for clean labels by Lamy et al. [10]. Our contribution is a single differentiable objective (Eq. (5)) that corrects both noise sources without clean validation data. The two corrections act on different quantities: the Forward Loss changes only the utility term through $\hat{y}_{fwd}$, while the MIRL penalty uses the uncorrected predictions $f_\theta(x)$ (Algorithm 1, steps 3–5), so correcting the labels does not alter the group statistics that MIRL inverts.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly state our theoretical claims (Theorem 1) and empirical claims (Section 7) in the abstract and introduction, and support them with mathematical proofs and extensive experiments.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in Section 7.5 (Boundary Conditions) and Section 8 (Future Work, Broader Societal Impacts): reliance on estimated noise rates, behavior in low-signal regimes, the approximation used for asymmetric noise (Section 5.7), and that the multi-class and Equal Opportunity extensions are derived but not evaluated.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The full set of assumptions for our noise model is stated in Section 5.1. The complete and correct proof for Theorem 1 is provided in Section 5.6.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide full details of our experimental setup, including datasets, noise injection protocols, and baseline implementations in Section 6.1, Section 7.3, and Appendix A.1.

Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: We have made our complete Python source code and Jupyter Notebook publicly available on GitHub (`https://github.com/gauravjain10/JRF`) to reproduce all experiments, tables, and figures.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Hyperparameters, network architectures, and training procedures are detailed in Section 7.3 and summarized in Table 3 of Appendix A.1.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report mean and standard deviation over 5 random seeds in Table 1 and Appendix A.2 (Tables 4 and 5); the sensitivity sweep (Table 6) uses 3 seeds.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The computational complexity is analyzed in Section 5.5. The experiments require negligible compute and were run on standard consumer hardware (a single CPU/GPU instance), taking less than 1 hour total.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics. Our work aims to improve algorithmic fairness and mitigate bias in machine learning systems.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: We explicitly discuss both the positive societal impacts (safe deployment in privacy-sensitive domains) and potential negative impacts (reliance on estimated noise rates) in Section 8.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: This paper proposes a training methodology and loss function, not a specific high-risk pre-trained model or scraped dataset.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We utilize standard, publicly available benchmarks (Adult, Bank, COMPAS) from the UCI Machine Learning Repository and OpenML, which are open for research purposes. Citations are provided in Section 6.1.

Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: This paper does not introduce new datasets or assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: This research does not involve human subjects and relies entirely on publicly available, anonymized datasets.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were not used as an important, original, or non-standard component of the core methodology. As declared in the submission form, they assisted with editing, drafting, implementing standard methods, and visualizing results; the author verified all content.

Guidelines:

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