Organization of the Supplement

- section A (§2):
  - motivating examples;
  - estimators for the noise under conditions stronger than assumption 1.
- section B (§3):
  - omitted proofs for section 3;
  - extension of results under conditions stronger than assumption 1;
  - extension of results under relaxation of assumption 2;
  - further experiments.
- Section C:
  - experiments on error rate balance with fairness-promoting algorithms.

A Extension for section 2

In this section, we use $m = m(x)$ and $m^* = m^*(x)$ to indicate $\mathbb{E}[Y|X = x]$ and $\mathbb{E}[Y^*|X = x]$ respectively. We also drop the dependency of γ on A and, if assumption 1 is used, on Y^*.

A.1 Who are the likely hidden recidivists?

In section §3 we have argued that the worst case bounds in our sensitivity analysis occur when the hidden recidivists are either all in the low-risk bin ($\alpha_1 = 0$) or all in the high-risk bin ($\alpha_0 = 0$). Here we present two thought examples reflecting on assumption 1. We show that, generally speaking, one can not rule out the "extreme" settings. Indeed, under assumption 1, the case $m \equiv 1$ is still possible.

· Example 1 · Suppose for instance that $X \in \{0, 1\}$ is a single binary covariate, $m^*(1) = 1, m^*(0) < 0.5$, and $\gamma(1) = 0.4, \gamma(0) = 0$. This gives $m(1) = 0.6$ and $m(0) = m^*_0 < 0.5 < 0.6 = m(1)$. If we set the classification threshold at $s_{HR} = 0.5$, we would classify everyone with $X = 1$ as high-risk and everyone with $X = 0$ as low-risk. By construction, we have $\gamma(0) = 0$, meaning that all recidivists with $X = 0$ are observed, whereas some fraction of recidivists with $X = 1$ are hidden. This in turn means that all hidden recidivists are classified as high-risk ($\alpha_0 = 0$). A similar construction can be used to produce a case where $\alpha_1 = 0$, which corresponds to all hidden recidivists being classified as low-risk.

· Example 2 · The first example is admittedly highly contrived and unlikely to reflect any real world scenario. To model a more plausible scenario, we consider a setup in which we have a single feature $X \sim Unif[0, 1]$, $m(x) = x$, and two forms for the likelihood of getting caught function:

$$\gamma^{Inc,b}(x) = 1 - (b + 1)x/(1 + bx),$$
$$\gamma^{Dec,b}(x) = 1 - (b + 1)(1 - x)/(1 + b(1 - x)).$$

The "Increasing" setting $\gamma^{Inc,b}$ is one where the likelihood of getting caught increases with the likelihood of reoffense $m^*(x)$, with the functional form of the relationship governed by the parameter b. The "Decreasing" setting has the likelihood of getting caught decreasing with the likelihood of reoffense. We equalize the proportion of high-risk and low-risk cases by thresholding $m(x)$ at its median value in each simulation. Figure 4 shows a plot of how the fraction of hidden recidivists that get classified as high-risk varies with b. Values larger than 0.5 on this plot can be interpreted as settings where $\alpha_1 > \alpha_0$; a value of 1, though never achieved, would correspond to the case $\alpha_0 = 0$. This suggests that, in general, the hidden recidivists are likely to be scattered across the range of the score S, and are thus unlikely to concentrate entirely in the extremes of S. In other words, the worst-case bounds presented in Section 3 are, unsurprisingly, likely to be overly conservative.

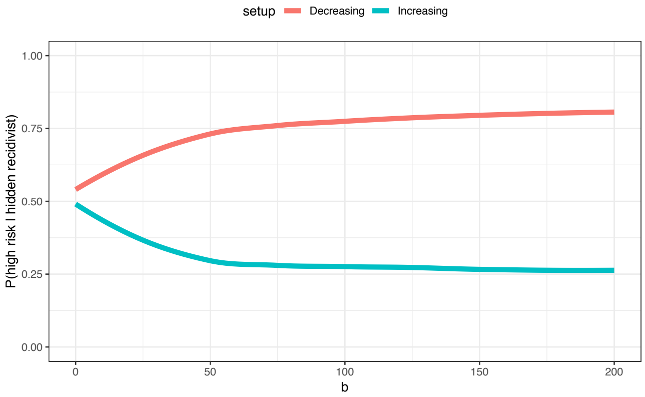


Figure 4: Proportion of hidden recidivists classified as high risk under different choices of γ.

A.2 Estimation of noise

In §2) we have argued that the assumption of constant noise is unrealistic in our setting. Indeed, in the introduction we cite the case of drug crimes (low-level offenses), where there appears to be an inconsistency in the number of arrests and users between the black and white populations; this fact might be attributed to *differential policing*. For other types of crimes we can imagine the effect of policing to be more similar across races. Although we suggest to account for more complex forms of the noise, one may wish to perform a sensitivity analysis under stronger assumptions on the noise process, e.g. assume the noise to be independent of the features conditionally on the observed labels. The case of constant noise has been intensively studied during the past two decades and it is fairly well understood. In this subsection we present a simple extension of this framework to account for noise constant within groups.

A.2.1 Estimation of one-sided label-dependent noise.

In the paper we work under the setup of assumption 1, that is of one-sided feature-dependent noise. Now, consider the following assumption.

Assumption 3. $Y \perp\!\!\!\perp X | Y^*$.

Under assumptions 1 and 3 we refer to the noise as *one-sided label-dependent*. Since the noise rate $\gamma(x, 1)$ is now constant, we drop the dependency on x and rewrite $\gamma = \gamma(x, 1)$.

We briefly describe three of the estimators for the noise rates commonly used in the literature. These estimators can be used for estimation of the noise rate in the setting of assumptions 1 and 3.

· Estimator 1 · The estimator proposed by (Elkan and Noto, 2008) relies on the following assumption.

Assumption 4. *(strong separability)* $m^*(x) \in \{0, 1\}$.

Then we have the following proposition.

Proposition A.1. *Under assumptions 1, 3, and 4, the following equality holds. For every* $y = 1$,

$$\gamma = 1 - m(x). \tag{11}$$

Proof of proposition A.1. Thanks to assumption 4, $y = 1$ implies $m^*(x) = 1$. Consequently we have $m(x) = (1 - \gamma)m^*(x) = 1 - \gamma$ for every $y = 1$. □

Estimators 2 and 3 rely on the following assumption.

Assumption 5. *(weak separability)* $\sup_x m^*(x) = 1$.

· Estimator 2 · The following is also described in (Elkan and Noto, 2008; Liu and Tao, 2016; Menon et al., 2015).

Proposition A.2. *Under assumptions 1, 3, and 5, the following equality holds.*

$$\gamma = 1 - \sup_x m(x) \tag{12}$$

Proof of proposition A.2. Recall the decomposition $m(x) = (1-\gamma)m^*(x)$. Then, thanks to assumption 5, we have

$$\sup_x m(x) = (1-\gamma)\sup_x m^*(x) = 1-\gamma \implies \gamma = 1 - \sup_x m(x).$$

□

Consequently the rate of convergence for the estimation of γ coincides with the one for $m(x)$.

· Estimator 3 · We define ρ, the *inverse noise rate*, as

$$\rho := \mathbb{E}[Y^*|Y=0] = \frac{\alpha}{1-\mathbb{E}[Y]} = \frac{\gamma}{1-\mathbb{E}[Y]}\mathbb{E}[Y^*] = \frac{\gamma}{1-\mathbb{E}[Y]}\frac{\mathbb{E}[Y]}{1-\gamma} = \frac{\gamma/(1-\gamma)}{(1-\mathbb{E}[Y])/\mathbb{E}[Y]}. \tag{13}$$

Note that γ identifies ρ, and vice versa. An estimator for ρ has been proposed by (Scott and Blanchard, 2009; Scott et al., 2013).
Let q_y^* and q_y denote the densities of X conditional on $Y^* = y$ and $Y = y$ respectively. Under assumptions 1, 3, and 5,

$$\rho = \frac{\nu(q_0, q_1^*)(1-\nu(q_1^*, q_0))}{1-\nu(q_1^*, q_0)\nu(q_0, q_1^*)} \tag{14}$$

where $\nu(q_1^*, q_0) = \inf_x q_1^*(x)/q_0(x)$ and $\nu(q_0, q_1^*) = \inf_x q_0(x)/q_1^*(x)$. ν corresponds to the left-derivative of the optimal ROC curve (Scott and Blanchard, 2009). The optimal ROC curve is given by any scorer that is a strictly monotone transformation of p (Clémençon et al., 2008). In (Scott and Blanchard, 2009) the estimator is recovered behind an assumption slightly weaker than assumption 5 that the authors call *irreducibility*; however, under this assumption, the convergence rate of the estimator is shown to be arbitrarily slow. (Scott, 2015) introduces an assumption equivalent to 5 that guarantees faster convergence rates.

It is clear that if assumption 5 does not hold, then the estimated noise rate is only upper bounded by $1-\sup_x m(x)$, and consequently $m(x) \leq m^*(x) \leq m(x) + \gamma$.

A.2.2 Estimation of one-sided race- and one-sided label-dependent noise.

In our setting it is more reasonable to consider a noise process that depends on the race membership; indeed, the original motivation of our work was a concern regarding *differential policing* across races. To simplify notation, let $m^a(x) := \mathbb{E}[Y|X=x, A=a]$; similarly, $\gamma^a := \mathbb{P}(Y=0|Y^*=1, A=a)$. We formulate the following assumption.

Assumption 6. $\sup_x m^{*a}(x) = 1 \;\; \forall a \in \{b, w\}$.

The unconditional version of assumption 6 is clearly assumption 5. The following proposition can be interpreted as a generalization of proposition A.2.

Proposition A.3. *Under assumptions 1, 3, and 6, the following equality holds.*

$$\gamma^a = 1 - \sup_x m^a(x) \;\; \forall a \in \{b, w\}. \tag{15}$$

Again, the convergence rate of the estimator of γ^a is identical to the one of the estimator of $m^a(x)$.

If race-specific classifiers are trained, then this framework inherits all the results from the label-dependent noise literature. Instead, if a unique classifier is trained, with race included in the feature set, then some of the results for model training and labels correction can be adapted to this setting.

We now estimate the values of γ^a on COMPAS data considering the setting of assumptions 1, 3, and 6. We fit one classifier for each race group and tune the parameters via cross-validation on the training set. We use extreme gradient boosted trees (xgboost) (Chen and Guestrin, 2016), logistic regression (glmnet), k-nearest neighbors (knn), and support vector machines (svm). The resulting scores are thresholded at 1/2 according to Bayes decision rule and the accuracy on the test set is approximately 66% for all models and both races. The results of the estimation for estimators 1 and 2, with corresponding standard deviations, are reported in Table 3. Not surprisingly, the noise parameter for the white population is higher than that for the black population across all models. This result is a consequence of violation of the assumptions – that are unlikely to hold in practice – and poor performance of the models.

Method	xgboost	glmnet	knn	svm
White/est (est 2)	0.13 (0.11)	0.18 (0.08)	0.12 (0.10)	0.15 (0.11)
White/est (est 1)	0.55 (0.02)	0.54 (0.02)	0.55 (0.01)	0.54 (0.01)
Black/est (est 2)	0.07 (0.06)	0.08 (0.05)	0.12 (0.08)	0.10 (0.08)
Black/est (est 1)	0.44 (0.02)	0.42 (0.02)	0.43 (0.02)	0.42 (0.02)

Table 3: The mean (standard deviation) values of γ^a estimated on 20 random train-test splits of COMPAS data are reported in the table for estimators (est) 1 and 2. The parameters of the models are tuned via cross validation. The feature set includes age, sex, count of juvenile felonies, count of juvenile misconduct, count of other juvenile charges, count of prior charges.

B Extension for Section 3

B.1 Omitted proofs

B.1.1 Error rates and predictive parity

Proof of proposition 3.1. Assume that $1 - FPR < FNR$. We now show by contradiction that $FNR \leq FNR^*$ and $FPR \geq FPR^*$ can not hold together. Indeed, the following two equivalences hold

$$FNR \leq FNR^* \iff FNR \leq \alpha_0/\alpha$$
$$FPR \geq FPR^* \iff 1 - FPR \geq \alpha_0/\alpha$$

thanks to corollary 3.2.1. It follows that $\alpha_0/\alpha \geq FNR > 1 - FPR \geq \alpha_0/\alpha$, which is a contradiction. The proof for the other case is analogous. □

Figure 5 provides a visual interpretation of the result.

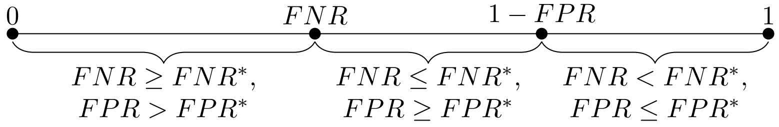

0
FNR
$1 - FPR$
1
$FNR \geq FNR^*$, $FPR > FPR^*$
$FNR \leq FNR^*$, $FPR \geq FPR^*$
$FNR < FNR^*$, $FPR \leq FPR^*$

Figure 5: Possible relationships between the true and observed error rates in the case of $1 - FPR > FNR$ as described in proposition 3.1.

Proof of theorem 3.2.1. Recall the following notation: $p_{ij} := \mathbb{P}(Y = i, \hat{Y} = j)$.

- *Proof of inequality* (1). $FPR = \mathbb{E}[\hat{Y}|Y = 0]$ can be rewritten as

$$\begin{aligned} &\mathbb{E}[\hat{Y}|Y^* = 0]\mathbb{P}(Y^* = 0|Y = 0) + \mathbb{E}[\hat{Y}|Y = 0, Y^* = 1]\mathbb{P}(Y^* = 1|Y = 0) \\ &= FPR^* \left(1 - \mathbb{E}[Y^*|Y = 0]\right) + \frac{\alpha_1}{\alpha}\mathbb{E}[Y^*|Y = 0] \end{aligned}$$

thanks to the law of total probability, Bayes theorem and assumption 1 in sequence. Therefore FPR is a convex combination of FPR^* and α_1/α. Rearranging the terms we obtain

$$FPR^* = \frac{FPR - \frac{\alpha_1}{\alpha}\mathbb{E}[Y^*|Y = 0]}{1 - \mathbb{E}[Y^*|Y = 0]} = \frac{p_{01} - \alpha_1}{p_{00} + p_{01} - \alpha}$$

For fixed $\alpha \leq \min\{p_{00}, p_{01}\}$, we obtain

$$\frac{p_{01} - \alpha}{p_{00} + p_{01} - \alpha} \leq FPR^* \leq \frac{p_{01}}{p_{00} + p_{01} - \alpha}.$$

- *Proof of inequality* (2). $FNR^* = \mathbb{P}(\hat{Y} = 0|Y^* = 1)$ can be rewritten as

$$\mathbb{P}(\hat{Y} = 0|Y = 1)\mathbb{P}(Y = 1|Y^* = 1) + \mathbb{P}(\hat{Y} = 0|Y = 0, Y^* = 1)\mathbb{P}(Y = 0|Y^* = 1).$$

Then we have

$$FNR^* = FNR\,\mathbb{E}[Y|Y^* = 1] + \frac{\alpha_0}{\alpha}(1 - \mathbb{E}[Y|Y^* = 1])$$

which is derived as above. The last derivation follows the same strategy as above.

- *Proof of inequality* (3). $PPV^* = \mathbb{E}[Y^*|\hat{Y} = 1]$ can be rewritten as

$$PPV + \mathbb{E}[Y^*(1 - Y)|\hat{Y} = 1] = PPV + \frac{\alpha_1}{p_{01} + p_{11}}.$$

Then, since the second term on the RHS is larger or equal to zero, the lower and upper bounds for PPV^* will be given by $\alpha_1 = 0$ and $\alpha_1 = \alpha$ respectively. □

Proof of corollary (3.2.1). Let us first prove equivalence (5).

$$FNR \geq FNR^{*} \iff \frac{p_{10}}{p_{10}+p_{11}} \geq \frac{p_{10}+\alpha_0}{p_{10}+p_{11}+\alpha} \iff FNR \geq \frac{\alpha_0}{\alpha}.$$

The proof of equivalence (4) for FPR is similar.

$$FPR \leq FPR^{*} \iff \frac{p_{01}}{p_{00}+p_{01}} \leq \frac{p_{01}-\alpha_1}{p_{00}+p_{01}-\alpha} \iff FPR \geq \frac{\alpha_1}{\alpha}.$$

□

Derivation of (6). Let us start with the case of FNR. If the condition in (5) holds, then we have

$$\frac{p_{10}}{p_{10}+p_{11}} \geq \frac{\alpha_0}{\alpha_0+\alpha_1} \iff \alpha_1 p_{10} \geq \alpha_0 p_{11} \iff \frac{\alpha_1/\alpha_0}{p_{11}/p_{10}} \geq 1$$

where we used Bayes theorem and law of total probability in sequence.
The odds ratio for FPR can be derived in a similar manner. For the equivalence in (4) to hold we need

$$\frac{p_{01}}{p_{00}+p_{01}} \geq \frac{\alpha_1}{\alpha_0+\alpha_1} \iff \alpha_0 p_{01} \geq \alpha_1 p_{00} \iff \frac{\alpha_0/\alpha_1}{p_{00}/p_{01}} \geq 1$$

where we used, again, Bayes theorem and law of total probability. □

B.1.2 Accuracy Equity

Proof of proposition 3.3. The Mann-Whitney U statistic can be computed according to

$$U_1 := R_1 - \frac{n_1(n_1+1)}{2} = \sum_{i:Y_i=1} r_i - \frac{n_1(n_1+1)}{2}$$

where r_i are the adjusted ranks. We can calculate the AUC of S as a classifier of Y^{*} from U_1 through the expression:

$$AUC^{*} = \frac{U_1}{n_1 n_0} = \frac{R_1}{n_1(n-n_1)} - \frac{n_1+1}{2(n-n_1)}$$

Now suppose that $\lceil \alpha n \rceil$ observations are unobserved recidivists. It is clear that the lower (upper) bound can be found by assuming $\lceil \alpha n \rceil$ observations corresponding to the lowest (highest) ranks such that $Y = 0$ to be recidivists; this provides the sharp bound in the proposition. This is in turn lower (upper) bounded by the case where the lowest (highest) $\lceil \alpha n \rceil$ ranks overall correspond to unobserved recidivists: for the lower bound, $R_1 = R_1 + 1 + \cdots + \alpha n = R_1 + (\alpha n + 1)\alpha n/2$, while for the upper bound, $R_1 = R_1 + (n - \alpha n + 1) + \cdots + n = R_1 + \alpha n(2n - \alpha n + 1)/2$. □

B.1.3 Calibration via logistic regression

Proof of proposition 3.4. For a fixed set proportion of hidden recidivists α, we aim to prove that the bounds for the coefficient of race are achieved in the settings $\alpha_1 = \alpha$ and $\alpha_0 = \alpha$.
Consider the random variables $\mathbf{X_i} = (X_{i,1}, X_{i,2}, X_{i,3})^T$ where $X_{i,1} = 1$, $X_{i,2} \in \mathbb{R}_{+}$, and $X_{i,3} = \mathbb{1}_w(A_i)$ with $A_i \in \{b, w\}$ $\forall i \in P = \{1, \ldots, n\}$. Let $W := \{i | i \in P \text{ and } x_{i,3} = 1\}$. Consider the n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$ such that $x_{i,2} \leq x_{j,2}$ for $1 \leq i \leq j \leq n$, that is the observations are ordered increasingly according to the realizations of $X_{i,2}$. Let $\boldsymbol{\beta}^{\dagger}$ be the MLE of the log-likelihood

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} y_i \log \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i) + (1 - y_i) \log(1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i)) \tag{16}$$

where

$$\sigma_{\boldsymbol{\beta}}(\mathbf{x}) := \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}}}.$$

Logistic regression aims at minimizing the negative log-likelihood in (16).
Consider two indices $l, h \in W$, $h > l$, such that $y_l = 1$ but $y_h = 0$. Now let $\{(y_i^*, \mathbf{x}_i)\}_{i=1}^n$ be such that $y_i^* = y_i \ \forall i \in P \setminus \{l, h\}$; $y_l^* = 0$ and $y_h^* = 1$. We are interested in the MLE $\boldsymbol{\beta}^{*\dagger}$ for $\ell(\boldsymbol{\beta}|\mathbf{y}^*)$. Consider a second-order Taylor expansion of $\ell(\boldsymbol{\beta}|\mathbf{y}^*)$ around $\boldsymbol{\beta}^\dagger$:

$$\ell(\boldsymbol{\beta}|\mathbf{y}^*) \approx \ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger)\nabla\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger)^T\nabla^2\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*)(\boldsymbol{\beta}^{*'} - \boldsymbol{\beta}^*).$$

Note that $\nabla\ell(\boldsymbol{\beta}|\mathbf{y}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\dagger} = (0, x_{h,2} - x_{l,2}, 0)^T$ since $\ell(\boldsymbol{\beta}|\mathbf{y}^*)$ can be rewritten as

$$\ell(\boldsymbol{\beta}|\mathbf{y}^*) = \ell(\boldsymbol{\beta}|\mathbf{y}) + \log\frac{\sigma_{\boldsymbol{\beta}}(\mathbf{x}_h)}{1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x}_h)} - \log\frac{\sigma_{\boldsymbol{\beta}}(\mathbf{x}_l)}{1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x}_l)}$$

where

$$\log\frac{\sigma_{\boldsymbol{\beta}}(\mathbf{x})}{1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x})} = \log\exp\{\boldsymbol{\beta}^T\mathbf{x}\} = \boldsymbol{\beta}^T\mathbf{x}$$

and thanks to the fact that the score evaluated at the MLE is zero. If we consider the problem of minimizing the negative log-likelihood, the Hessian is positive definite, and consequently its determinant is positive. We are interested in the direction of the search for $\boldsymbol{\beta}^{*\dagger}$. The minimizer of the Taylor expansion above for the negative log-likelihood with respect to $\{(y_i^*, \mathbf{x}_i)\}_{i=1}^n$ is $\boldsymbol{\beta}^+ = \boldsymbol{\beta}^\dagger - \left[\nabla^2\left(-\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*)\right)\right]^{-1}\nabla(-\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*))$. The Hessian is given by $\mathbf{X}^T\mathbf{D}\mathbf{X}$ where $\mathbf{D}_{ii} = s_i := \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i)(1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i))$ for $i = 1, \dots, n$. Therefore we have

$$\nabla^2\left(-\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*)\right) = \mathbf{X}^T\mathbf{D}\mathbf{X} = \begin{bmatrix} \sum_{i\in P} s_i & \sum_{i\in P} s_i x_{i,2} & \sum_{i\in W} s_i \\ \sum_{i\in P} x_{i,2}s_i & \sum_{i\in P} s_i x_{i,2}^2 & \sum_{i\in W} s_i x_{i,2} \\ \sum_{i\in W} s_i & \sum_{i\in W} x_{i,2}s_i & \sum_{i\in W} s_i \end{bmatrix}$$

Since the gradient of $-\ell(\boldsymbol{\beta}^\dagger|\mathbf{y}^*)$ is $(0, x_{l,2} - x_{h,2}, 0)$, we are only interested in the second column of the inverse of the Hessian. Through some algebra to invert the Hessian, we obtain that $\beta_k^+ \leq \beta_k^\dagger$ for $k = 1$ and $\beta_k^+ \geq \beta_k^\dagger$ for $k = 2, 3$ if the following respective conditions hold:

1. $\sum_{i\in P\setminus W} s_i x_{i,2} \geq 0$ for k=1;
2. $\sum_{i\in P\setminus W} s_i \geq 0$ for k=2;
3. $(\sum_{i\in W} s_i)(\sum_{i\in P} x_{i,2}) - (\sum_{i\in W} s_i x_{i,2})(\sum_{i\in P} s_i) \geq 0$ for k=3;

where $s_i = \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i)(1 - \sigma_{\boldsymbol{\beta}}(\mathbf{x}_i))$. Notice that condition (2) will always be verified, and condition (1) as well if $X_2 \in \mathbb{R}_+$, as in our case. Condition (3) needs to be verified case by case. It follows that, if condition (3) holds for any choice of $h > l$, then the coefficient of race is a nondecreasing function of the index. □

For varying α, one can prove the inequality using a similar approach. For a model with $X = (X_1, X_2)$ the proof is straightforward using a first-order Taylor expansion. With the inclusion of an additional covariate X_4, the gradient becomes $(0, x_{h,2} - x_{l,2}, 0, x_{h,4} - x_{l,4})^T$ and the inversion of the Hessian is not straightforward.

B.1.4 Optimization for sensitivity analysis of chi-squared conditional independence test

We recall the test statistic of the chi-squared test;

$$T(h) = \sum_{k=1}^{|S|} \sum_{\substack{a\in\{b,w\} \\ y\in\{0,1\}}} \frac{\left(O_{ay}^{(k)} - E_{ay}^{(k)}\right)^2}{E_{ay}^{(k)}}, \tag{17}$$

The statistic is a function of the hidden recidivist counts $h = (h_1, \dots, h_{|S|})$. Expected counts are estimated from the data assuming the null hypothesis $Y^* \perp\!\!\!\perp A \mid S$ is true. These quantities evaluate to

$$\begin{aligned} O_{ay}^{(k)} &= n_{ay}^{(k)} + h_k \mathbb{1}_{a=w}(2y - 1), \text{ and} \\ E_{ay}^{(k)} &= \left(n_{wy}^{(k)} + n_{by}^{(k)} + (2y - 1)h_k\right)\left(n_{a0}^{(k)} + n_{a1}^{(k)}\right)/n^{(k)}. \end{aligned}$$

The key observation is that, when viewed as a function of h_k, the numerator terms $(O_{ay}^{(k)} - E_{ay}^{(k)})^2$ are convex quadratics in h_k, and the denominator terms $E_{ay}^{(k)}$ are linear functions in h_k that are constrained to be positive. Thus each inner summand of equation (17) is a quadratic-over-linear function, which is strongly convex (Boyd and Vandenberghe, 2004). Furthermore, since the sum of strongly convex functions is strongly convex, we can conclude that the test statistic T as a function of h has the form

$$T(h) = \sum_{k=1}^{|S|} f_k(h_k), \tag{18}$$

where each f_k is a strongly convex function. This observation is important in our discussion of optimizing the test statistic subject to constraints on the hidden recidivist population.

Now, we want to maximize the test statistic (17) over h_k, subject to $\sum_k h_k \leq N_h$.' Note that each term is strongly convex in h_k, so the optimum over h_k for $0 \leq h_k \leq C$ will always be achieved at either $h_k = 0$ or $h_k = \min\{C, n_{w0}^{(k)}\}$. Because the objective is separable, we just take these terms in order of decreasing value in a simple greedy search:

```
Require: $N_h$                                                        ▷ Move limit
Require: $T_k(h_k)$                              ▷ Terms of (13) corresponding to k
Require: $n_{w0}^{(k)}$                                             ▷ See Section 3.4
  $B \leftarrow N_h$
  $h_k \leftarrow 0$, $k = 1, \ldots, K$
  while $B > 0$ do
      for $k \leftarrow 1$ to $K$ do
          $r[k] \leftarrow \max(0, T_k(\min(B, n_{w0}^{(n)} - h_k)) - T_k(0))$
      end for
      if $\max(r) \leq 0$ then
          Break while loop
      end if
      $i \leftarrow \text{argmax}(r)$                        ▷ Select greatest improvement
      $h_k \leftarrow \min(B, n_{w0}^{(n)})$
      $B \leftarrow B - \min(B, n_{w0}^{(n)})$
  end while
```

B.2 Extension to one-sided label-dependent noise

Recall from §A.2.1 that $\rho := \mathbb{E}[Y^*|Y = 0]$.

B.2.1 Error rate balance and predictive parity.

The following result can be read as a corollary of theorem 3.2.1. The decompositions of FPR^* and FNR^* have already been derived in (Jain et al., 2017; Scott et al., 2013; Menon et al., 2015).

Corollary B.0.1. *Under assumptions 1 and 3,*

$$FPR^* = \frac{FPR - \rho(1 - FNR^*)}{1 - \rho} \tag{19}$$

$$FNR^* = FNR \tag{20}$$

$$PPV^* = \frac{PPV}{1 - \gamma} \tag{21}$$

Proof of corollary B.0.1.

- *Proof of equation* (19). Consider the decomposition

$$FPR = (1 - \rho)\mathbb{E}[\hat{Y}|Y^* = 0, Y = 0] + \rho\mathbb{E}[\hat{Y}|Y^* = 1, Y = 0]$$

derived in the proof of theorem 3.2. Then,

$$\mathbb{E}[\hat{Y}|Y^* = 0, Y = 0] = \mathbb{E}[\hat{Y}|Y^* = 0] = FPR^*$$

and

$$\mathbb{E}[\hat{Y}|Y^* = 1, Y = 0] = \mathbb{E}[\hat{Y}|Y^* = 1] := 1 - FNR^*.$$

The result follows.

- *Proof of equation* (20). For FNR^* we have

$$\begin{aligned} 1 - FNR = \mathbb{E}[\hat{Y}|Y = 1] &= \sum_{y=0}^{1} \mathbb{P}(\hat{Y} = 1, Y^* = y|Y = 1) \\ &= \mathbb{P}(\hat{Y} = 1, Y^* = 1|Y = 1) = \mathbb{E}[\hat{Y}|Y^* = 1, Y = 1] \\ &= \mathbb{E}[\hat{Y}|Y^* = 1] = 1 - FNR^* \end{aligned}$$

and therefore $FNR = FNR^*$.

- *Proof of equation* (21). For PPV^*, similarly to the previous proofs,

$$PPV = \mathbb{E}[Y|Y^* = 1, \hat{Y} = 1]\mathbb{E}[Y^* = 1|\hat{Y} = 1] = (1 - \gamma)PPV^*.$$

□

B.2.2 Accuracy Equity

Proposition B.1. *Under assumptions 1 and 3,*

$$AUC = (1 - \rho)AUC^* + \rho/2. \tag{22}$$

Proof of proposition B.1. The resut follows from corollary 3 in (Menon et al., 2015) for the two-sided label-dependent noise setting considering $\beta = \rho$ and $\alpha = 0$. □

B.2.3 Calibration via logistic regression.

Thanks to assumptions 1 and 3, we have

$$\mathbb{E}[Y|S = s, A = w] = (1 - \gamma)\mathbb{E}[Y^*|S = s, A = w] \tag{23}$$

for all values of s, hence calibration properties can be easily checked.
In this context, the sensitivity analysis for calibration described in the paper (§3) still applies. The "extreme" settings $\alpha_0 = \alpha$ and $\alpha_1 = \alpha$ are ruled out only in expectation. In fact, when selecting hidden recidivists from the pool $\{Y = 0\}$ under condition 3,

$$\#\{Y \neq Y^*, \hat{Y} = 0\} \sim \text{Hypergeometric}(\#\{Y = 0\}, \#\{Y = 0, \hat{Y} = 0\}, \#\{Y \neq Y^*\}).$$

Therefore all hidden recidivists might still happen to be in the either lowest or highest risk bins.

However, one might want to correct the model in the training phase. For this purpose, several techniques inherited from the literature on label-dependent noise can be applied. For instance, the following two-step technique can be used: (1st step) training of *any* classifier and estimation of the noise rate, (2nd step) training of a logistic regression using the methods of unbiased estimators or of label-dependent costs proposed by (Natarajan et al., 2013). We provide below a quick overview of the two methods; further details can be found in (Natarajan et al., 2013).

· Method of unbiased estimators · For a scorer f and a bounded loss function $\ell(f(x), y)$,

$$\mathbb{E}_{Y|Y^*=y^*}\left[\tilde{\ell}(f(x), Y)\right] = \ell(f(x), y^*)$$

where

$$\tilde{\ell}(f(x),0) = \ell(f(x),0), \text{ and } \tilde{\ell}(f(x),1) = \frac{\ell(f(x),1) - \gamma\ell(f(x),0)}{1-\gamma} = \frac{\ell(f(x),1)(1+\gamma) - \gamma}{1-\gamma}.$$

The last equality is thanks to the fact that for the sigmoid loss $\ell(f(x), y) = (1 + \mathrm{e}^{-f(x)})^{y-1}(1 + \mathrm{e}^{f(x)})^{-y}$ we have $\ell(f(x), 0) + \ell(f(x), 1) = 1$. Therefore the optimization problem on noisy labels can be solved using the loss $\tilde{\ell}$ instead of ℓ.

· Method of label-dependent costs · For any classfier h, the Bayes classifier for the $0-1$ β-weighted loss function for Y is

$$(1-\beta)\mathbb{1}_{Y=1}\mathbb{1}_{h(X)=0} + \beta\mathbb{1}_{Y=0}\mathbb{1}_{h(X)=1}, \qquad \beta = (1-\gamma)/2.$$

This also corresponds to be the Bayes classifier for the minimization of the $0-1$ loss function for Y. (Natarajan et al., 2013) show that the use of the sigmoid loss ℓ as surrogate, that is the minimization of

$$(1-\beta)\mathbb{1}_{Y=1}\ell(f(X),1) + \beta\mathbb{1}_{Y=0}\ell(f(X),0),$$

ensures convergence of the $0-1$ β-weighted loss function.

B.3 Extension to noise in both groups

In this subsection we show that most of the results in our methodology extend to the case of noise in both groups without further proofs. Indeed, the results relative to error rates and AUC have been derived conditioning on the race attribute A. The proof for logistic regression can be easily adapted to take into account the new setting.

Let $\alpha_i^a := \mathbb{P}(Y=0, Y^*=1, \hat{Y}=\hat{y}|A=a)$ indicate the proportion of hidden recidivists in the low ($\hat{y}=0$) and high ($\hat{y}=1$) risk groups for the black ($a=b$) and white ($a=b$) populations. Let $\alpha^a := \alpha_0^a + \alpha_1^a$ be the total proportion of hidden recidivism in the population with race a.

Error rates and predictive parity. The bounds in proposition 3.1 and in theorem 3.2.1 have been obtained conditioning on the race attribute, that is M^{*a} only depends on $(M^a, \alpha_0^a, \alpha_1^a)$. This means that the sensitivity analysis on the metrics of an individual race group does not depend on the noise present in other groups. Consequently the results of proposition 3.1, theorem 3.2, and corollary 3.2.1 translate onto this setting without further proofs.
In the paper we also show that, in absence of noise for the black population, $FNR^w - FNR^{*b} \geq FNR^{*w} - FNR^{*b}$ whenever $FNR^w \geq \alpha_0^w/\alpha^w$ thanks to corollary 3.2.1. It is clear that $FNR^w - FNR^b \geq FNR^{*w} - FNR^{*b}$ will hold if we assume $FNR^b \leq \alpha_0^b/\alpha^b$ and $FNR^w \geq \alpha_0^w/\alpha^w$; however, it is unlikely – but not impossible – that the inequality holds in different directions for the two populations. Therefore an interesting question is what assumptions on $(\alpha_0^w, \alpha_1^w, \alpha_0^b, \alpha_1^w)$ are needed to conclude $FNR^w - FNR^b \geq FNR^{*w} - FNR^{*b}$. Through some algebra we can retrieve the following decomposition.

$$FNR^w - FNR^b = \frac{FNR^{*w}}{\mathbb{E}[Y|Y^*=1, A=w]} - \frac{FNR^{*b}}{\mathbb{E}[Y|Y^*=1, A=b]} + \frac{\alpha_0^b}{\mathbb{E}[Y|A=b]} - \frac{\alpha_0^w}{\mathbb{E}[Y|A=w]}.$$

The *differential policing* assumption would suggest that $\mathbb{E}[Y|A=w, Y^*=1] \leq \mathbb{E}[Y|A=b, Y^*=1]$ therefore we can lower bound the first two terms by $FPR^{*w} - FNR^{*b}$. There only remains to show that the last two terms are larger or equal to zero. However, this is not always the case. Indeed,

$$\frac{\alpha_0^b}{\mathbb{E}[Y|A=b]} \geq \frac{\alpha_0^w}{\mathbb{E}[Y|A=w]} \iff \frac{\alpha_0^b}{\alpha_0^w} \geq \frac{\mathbb{E}[Y|A=b]}{\mathbb{E}[Y|A=w]}.$$

In the COMPAS data we have seen that the RHS is larger than one, but we would intuitively expect that LHS to be smaller than one. Therefore we conclude that in order to make inference on the sign of $FNR^{*w} - FNR^{*b}$, explicit assumptions on the magnitude of the noise parameters need to be formulated, that is $\alpha_0^w, \alpha_1^w, \alpha_0^b$, and α_1^w need to be bounded. We do not present the computations for FPR, but the inequality has a similar interpretation. Finally, note that theorem 3.2 and corollary 3.2.1 can be rewritten in terms of unconditional statements, that is on the entire population. This is the typical setup in the literature when there is no specific interest in the conditional metrics.

Accuracy Equity. As in the case of error rates, the statement in proposition 3.3 holds conditionally on the protected attribute. Consequently no further extension is needed.
Again, we remark that the statement of the proposition holds also unconditionally, or, in general, conditionally on any subset of the feature space.

Calibration via logistic regression. We provide only a high-level idea for the extension of the proof of proposition 3.4. The gradient of the log-likelihood now is $\nabla\ell(\boldsymbol{\beta}|\mathbf{y}^*) = (0, x_{h,2} - x_{l,2}, x_{h,3} - x_{l,3})$ where $x_{h,3} - x_{l,3}$ is equal to 0 if $x_{h,3} = x_{l,3}$, that is if the hidden recidivist does not switch race. Consequently, for a fixed configuration of hidden recidivists in one population, the bounds for the coefficients will still be achieved by the hidden recidivists taking the extreme scores. Therefore one can show that, considering a pair of hidden recidivists of different races with scores either both lower or larger than the current ones, the bounds for the coefficients are achieved in the extreme settings over the entire population.

B.4 Further experiments for calibration via logistic regression

Figure 6 shows the two-dimensional bounds (red lines) of the coefficients of S and A for varying α, α_0, and α_1 as described by proposition 3.4. Although the analytical bounds for the coefficients for fixed α are wide, we find empirically that no matter the indexes of the hidden recidivists, the coefficients at a given α always lie on the diagonal (black lines) connecting the lowest and highest bounds that we find. Moreover, as previously argued, assuming label-dependent noise does not drastically change the coefficient of S but only the one of A, as shown by the coefficients obtained "randomly" sampling hidden recidivists from the observations with $Y = 0$ (orange lines).

We also check calibration for race- and label-dependent noise, i.e. under assumption 6, in the COMPAS data. The methodology follows the method of label-dependent costs described in §B.2.3. The procedure for the estimation of the race-specific noise rates has been described in §A.2.1; we use extreme gradient boosted trees for this step (Chen and Guestrin, 2016). We resample the observations from the data set according to the weights β described in §B.2.3; this is done separately within each of the two races. We then fit a logistic regression $Y \sim S + A$ on the resulting data set and check calibration via a Wald test. As in the observed data, the coefficient for $A = w$ is not statistically significant at an α-level of 0.01.

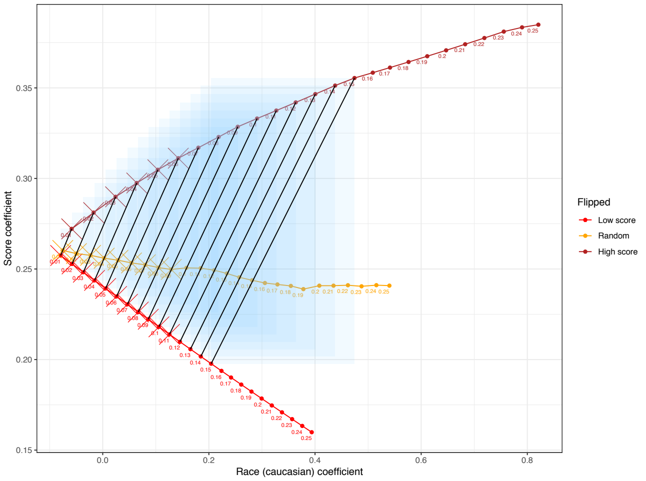


Figure 6: Outer red curves show the bounds guaranteed by proposition 3.4 for different choices of α (which are labeled on the curves). Blue rectangles correspond to analytical bounds, though empirically we can verify that the coefficients must in fact lie on the dashed diagonals. Orange curve shows coefficients under a "random" mechanism where we assume each observation with $Y = 0$ was equally likely to be a hidden recidivist. Points marked with an X correspond to values of α^w where the p-value of β_w is *not* statistically significant. These are values of α^w where we would conclude that S is well calibrated with respect to A as a predictor of Y^*.

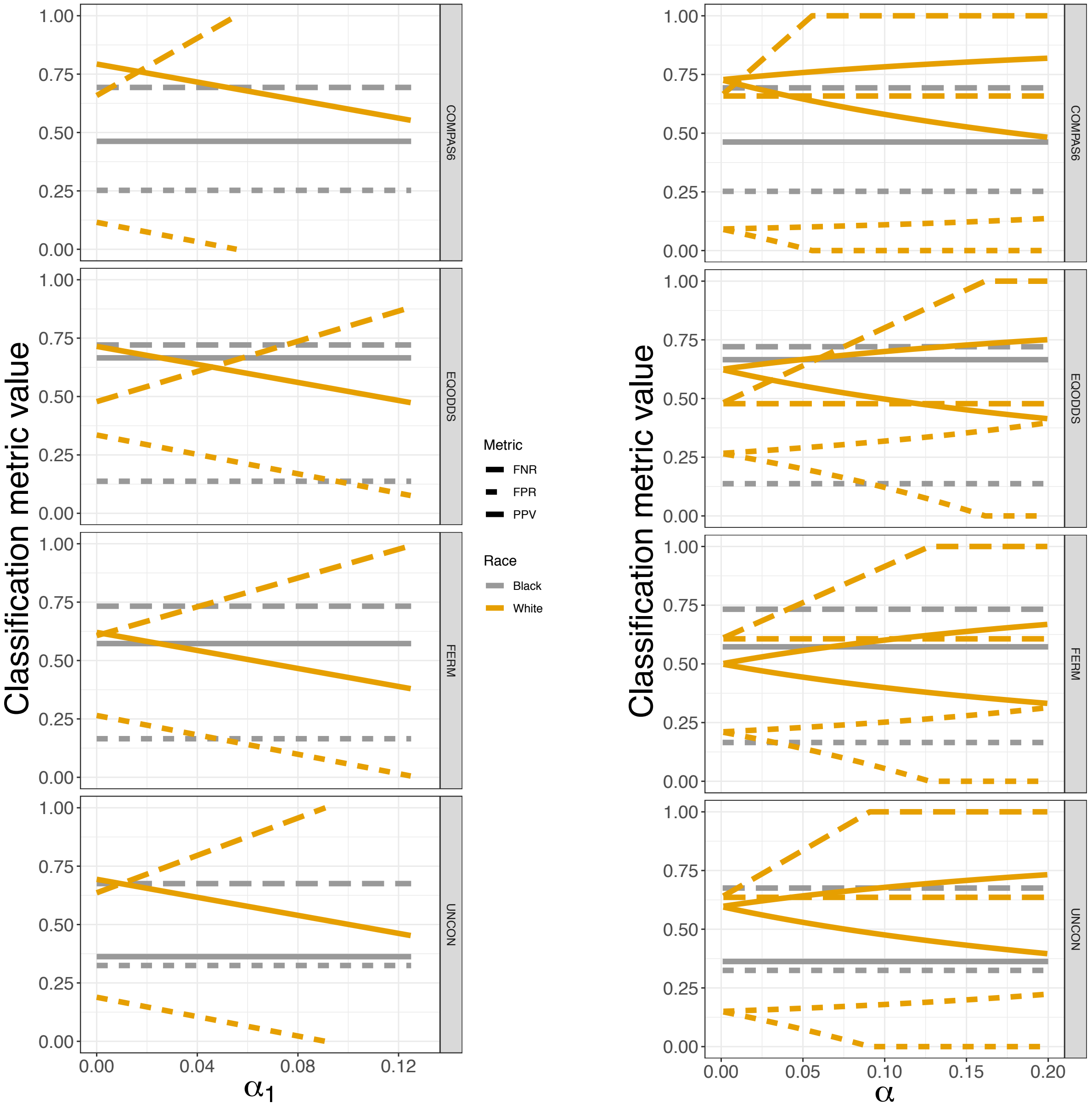
C Error rate balance with fairness-promoting algorithms

Through our methodology, we evaluate the effects of label noise on the error metrics of the predictions of the following four algorithms on the COMPAS data set. We split the data into 70% and 30% for training and testing respectively, stratifying for race. As feature set, we consider race, sex, age, number of juvenile felonies, misdemeanors, and other charges, count of prior arrests, degree of charge to predict two-year rearrest.

- (FERM) We use the methodology proposed by (Donini et al., 2018), training SVM's with linear kernel to produce a classifier that approximately satisfies equal opportunity.
- (EQODDS) We train a logistic regression and then use the methodology described in (Hardt et al., 2016) to obtain a classifier that satisfies equal opportunity.
- (COMPAS6) We threshold the COMPAS decile score at 6 (i.e. $\hat{Y} = \mathbb{1}(S > 6)$), instead of 4.
- (UNCON) We train a logistic regression.

We chose the thresholds for (COMPAS6) and logistic regression models such that the proportion of defendants predicted to be high risk was equal across all methods, i.e. around 30%.

The bounds for the error metrics for the predictions of the four classifiers as functions of the noise are shown in Figure 7. For varying α, we observe that the classifiers in (COMPAS6) and (UNCON) do not satisfy error rate balance on the observed labels. Due to the large differences in error rates, equality of the metrics of these models cannot be achieved by any configuration of the noise for $\alpha \leq 0.2$. Differently, equality is possible for (EQODDS) and (FERM) for α larger than 0.08, well below the level of 0.12 necessary to equalize reoffense rates. When the noise is fixed at $\alpha = 0.12$, we observe a similar pattern. Despite the unavoidable degree of uncertainty, (FERM) comes close to achieving parity: for $\alpha_1 = 0.04$, the metrics of (FERM) are approximately equal across populations. These results suggest that the two presented fairness-promoting methods perform better than unconstrained methods under label noise.



(a) $\alpha = 0.12$ to equalize reoffense rates across groups.

(b) Bounds as described in Theorem 3.2 in terms of α.

Figure 7: Analysis of predictive parity and error rate balance for COMPAS across different TVB scenarios for four different algorithms, as described in the text. Orange lines show values of FPR^w, FNR^w, and PPV^w. Grey lines show corresponding values for the black population.