Experiments

In this section we evaluate the performance of AFA algorithm against i.i.d estimator (See Appendix A.2 for details). Since i.i.d estimators depend directly on the property of interest, each estimation requires a different sampling strategy. Consequently, i.i.d estimators require a different interaction strategy with the black-box oracle. In contrast, AFA algorithm interacts once with the black box oracle and output (ϵ, δ)-PAC estimations for all the distributional properties of interest. We a commonly used dataset in the fairness literature: COMPAS dataset. We train Logistic Regression model. The model will be later accessible to the AFA algorithm with limited budget of queries (black-box setting). Each property true value is computed from the dataset.

COMPAS Dataset

The input space dimension is $n = 13$. The sensitive attribute for group fairness is the binary feature that splits the feature space into Caucasian and non-Caucasian. The evaluation for robustness and individual fairness depend on the perturbation parameters, table 1 summarizes the range of values of perturbation parameter ρ, while fixing $l = 10$.

Robustness and Individual Fairness: The perturbation parameter ρ can vary on $(-1, 1)$. Given the protocols of perturbation, small ρ values correspond to high perturbations. $\rho = -1$ corresponds to the highest possible perturbation of the feature space. Small ρ values correspond then to a shift in the distribution $\mathcal{D}$. Since AFA algorithm queries from the streaming distribution, it is expected to perform poorly to high range of perturbations.

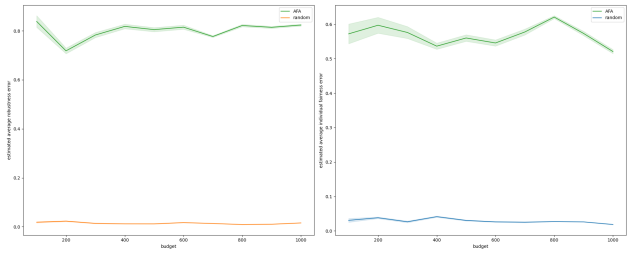


Figure 1. Comparison of robustness and inidividual fairness estimation error over query budgets: AFA auditor Versus random estimator for $\rho = -1$.

As anticipated, the poor performance of AFA algorithm shown in figure 1 (Experiment 1 in table 1) aligns with our expectations. This is due the the shift in the distribution by the perturbation protocol. In the range $\rho \in (-1, 0)$, Consistent with our earlier anticipation, Figures 2 and 3 show that AFA maintains a bad performance against random estimator.

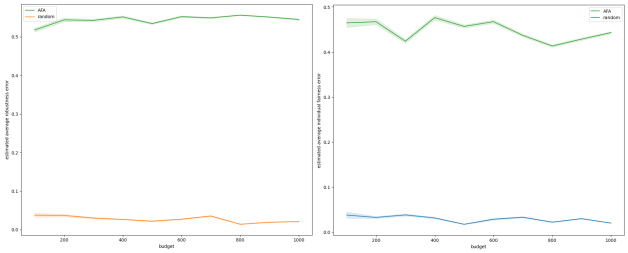


Figure 2. Comparison of robustness and individual fairness estimation error over query budgets: AFA auditor Versus random estimator for $\rho = -0.5$.

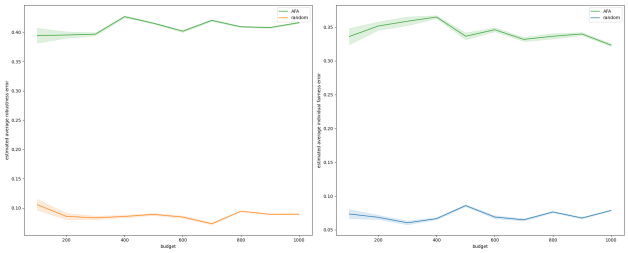
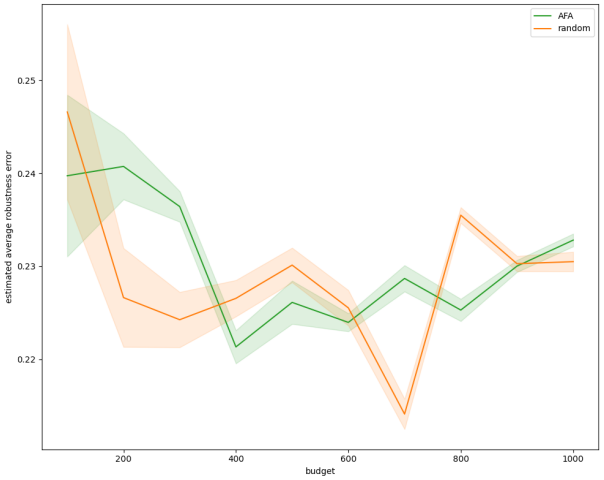


Figure 3. Comparison of robustness and individual fairness estimation error over query budgets: AFA auditor Versus random estimator for $\rho = -0.25$.

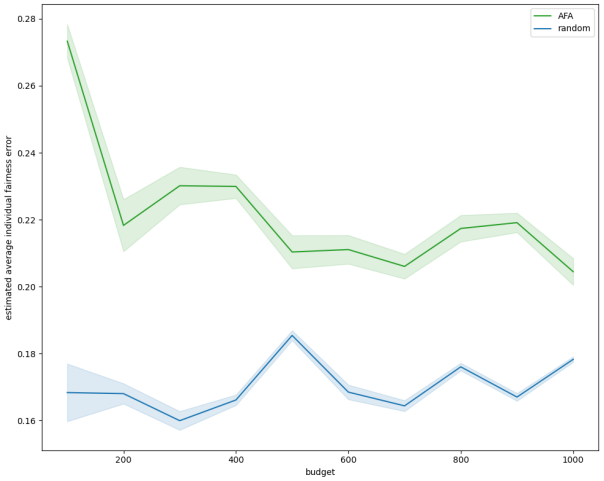
Figure 4 corresponding to $\rho = 0$ shows the competitiveness of AFA with the random estimator. We observe that $\rho = 0$ is an inflection point on the performance of AFA algorithm. Negative values of ρ correspond to very high perturbations that AFA fails to capture. On the other hand, positive and reasonably small ρ values correspond to plausible perturbations. As figure 4 shows, AFA auditor becomes very competitive for reasonable perturbation parameters.

ρ parameter	AFA μ_{Rob} error	random μ_{Rob} error	AFA μ_{IFair} error	random μ_{IFair} error
-1.00	0.823	$\mathbf{1.5 \times 10^{-2}}$	0.520	$\mathbf{1.8 \times 10^{-2}}$
-0.50	0.545	$\mathbf{2.0 \times 10^{-2}}$	0.442	$\mathbf{1.9 \times 10^{-2}}$
-0.25	0.416	$\mathbf{8.9 \times 10^{-2}}$	0.323	$\mathbf{7.8 \times 10^{-2}}$
0.00	0.232	**0.230**	0.204	**0.178**
0.30	**0.139**	0.299	**0.092**	0.248
0.35	**0.112**	0.309	**0.086**	0.273
0.40	**0.078**	0.333	**0.047**	0.309

Table 1. A summary of theoretical results: This table summarizes the expression of the estimation for each property with query complexity and computational complexity.



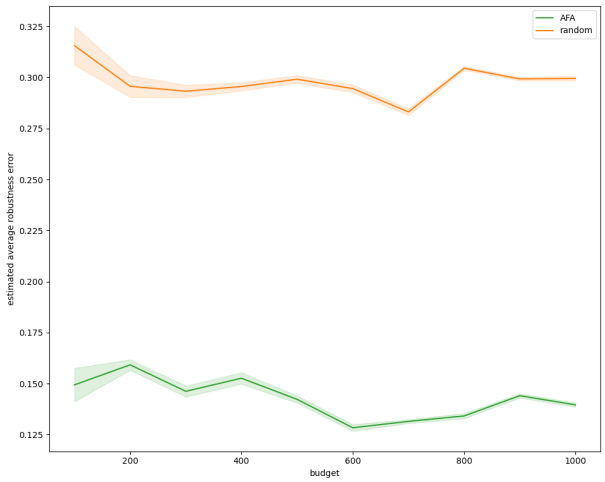
(a) Robustness



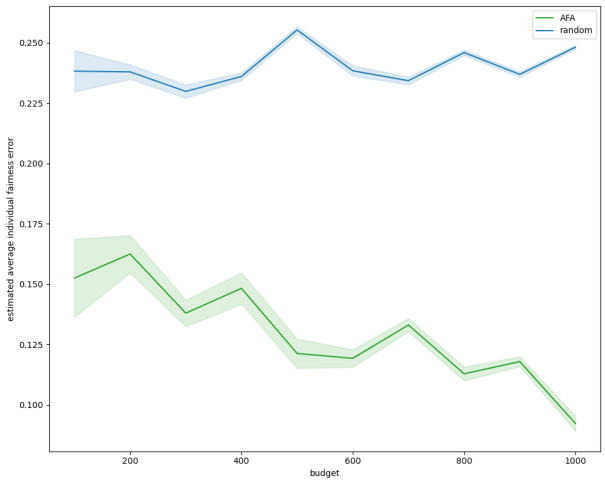
(b) Individual fairness

Figure 4. Comparison of estimation error over query budgets on COMPAS dataset: AFA auditor Versus Random estimator for $\rho = 0$.

For positive values of ρ, and As we increment the perturbation parameter , the performance of our proposed algorithm consistently outperforms the baseline in both robustness and individual fairness, maintaining a consistent level of effectiveness as shown in figures 5, 6 and 7.

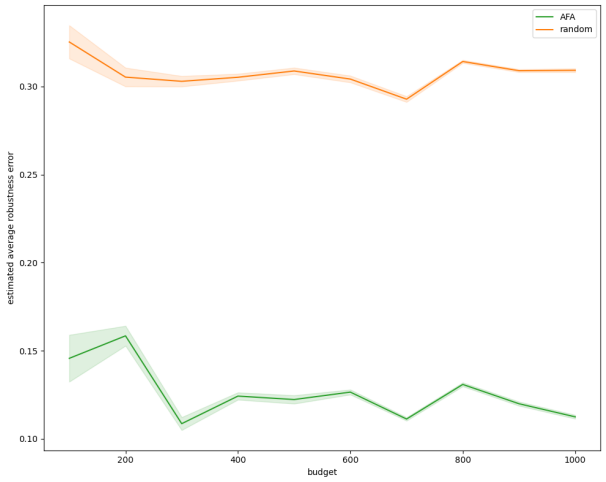


(a) Robustness

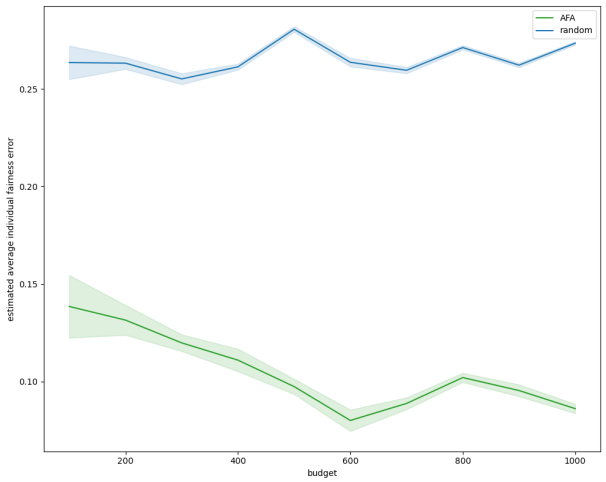


(b) Individual fairness

Figure 5. Comparison of estimation error over query budgets: AFA auditor Versus random estimator for $\rho = 0.3$.

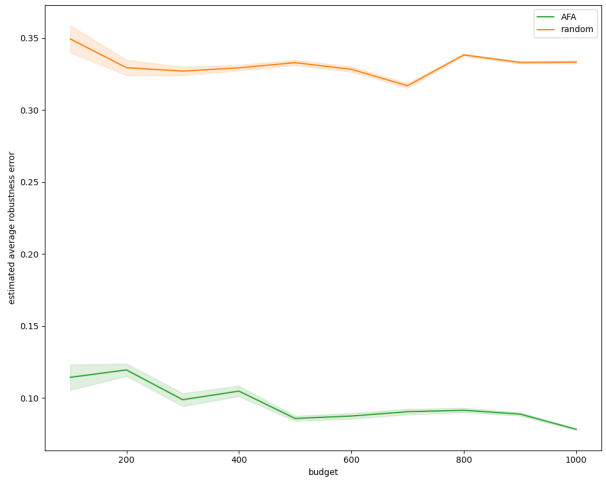


(a) Robustness

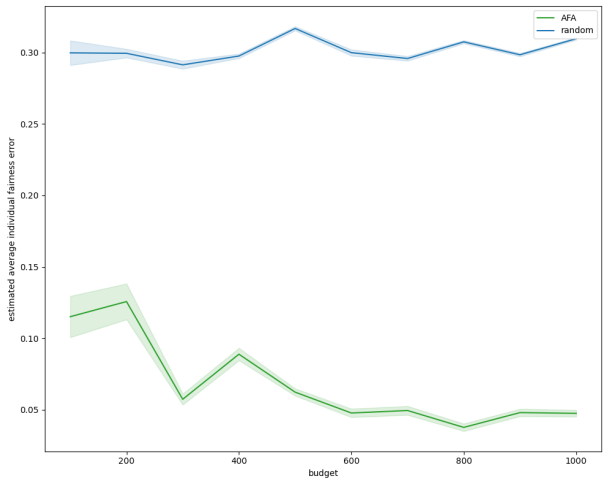


(b) Individual fairness

Figure 6. Comparison of estimation error over query budgets: AFA auditor Versus Random estimator for $\rho = 0.35$.



(a) Robustness



(b) Individual fairness

Figure 7. Comparison of estimation error over query budgets: AFA auditor Versus Random estimator for $\rho = 0.40$.

We have seen that AFA algorithm outperforms random estimator for small perturbation, which shows the high sensitivity of AFA to distribution shift. In contrast, AFA exhibits robustness in the face of variations in the parameter l as shown in appendix B.

Remark: We observe an interesting phenomenon that occurs at the phase transition around $\rho = 0$, where both AFA and random estimator performances become reversal.

Group Fairness: We evaluated AFA auditor on group fairness and compared its performance to the random estimator. AFA achieves high performance and converges rapidly to the true value of group fairness.

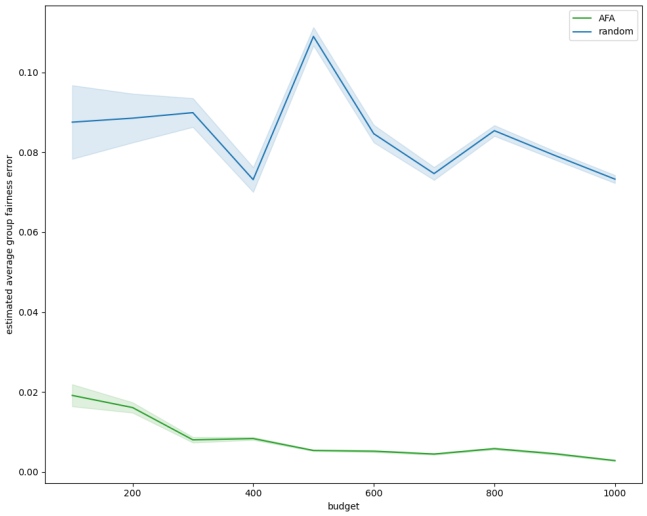


Figure 8. Comparison of group fairness estimation error over query budgets: AFA auditor Versus random estimator.

Remark: Specifically, for group fairness, we can optimize Goldreich-Levin to restrict the search of significant Fourier coefficients to buckets containing the sensitive attribute, resulting in exponential gains in computational complexity.

A. Experimental details

A.1. True properties computation

Let P denotes the whole dataset points. We evaluate the estimators performance according to true value of the properties given by the pool P. For fixed perturbation parameters ρ and l, we express the true value of the distributional properties.

Robustness: Robustness property is:

$$\mu_{\mathrm{Rob}}(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} [h(\mathbf{x}) \neq h(\mathbf{y})]$$

Since the true property is computed from P, we have:

$$\mu_{\mathrm{Rob}}(h) = \frac{1}{|P|} \sum_{\substack{\mathbf{x}\in P\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Individual Fairness: Similarly, individual fairness is:

$$\mu_{\mathrm{IFair}}(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}} [h(\mathbf{x}) \neq h(\mathbf{y})]$$

Since the true individual fairness is computed from P, we have:

$$\mu_{\mathrm{IFair}}(h) = \frac{1}{|P|} \sum_{\substack{\mathbf{x}\in P\\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Group Fairness: Let A denotes the protected attribute that partition the pool P: $P = P^+ \cup P^-$, where P^+ denotes the first protected group and P^- the second protected group. Group Fairness (with demographic parity measure) is defined as:

$$\mu_{\mathrm{GFair}}(h) = \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = 1] - \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}}[h(\mathbf{x}) = 1 | x_A = -1]$$

The true group fairness is:

$$\mu_{\mathrm{GFair}}(h) = \frac{1}{|P^+|} \sum_{\mathbf{x}\in P^+} \mathbb{1}_{h(\mathbf{x})=1} - \frac{1}{|P^-|} \sum_{\mathbf{x}\in P^-} \mathbb{1}_{h(\mathbf{x})=1}$$

A.2. Random estimators (I.I.D estimators):

Random estimators use i.i.d sampling in order to estimate each distributional property. We'll note that group fairness estimation requires a different sampling strategy and interaction with the black-box oracle of h.

The true robustness is defined as:

$$\mu_{\mathrm{Rob}}(h) = \mathop{\mathbb{P}}_{\substack{\mathbf{x}\sim\mathcal{D}\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} [h(\mathbf{x}) \neq h(\mathbf{y})]$$

Random estimator samples i.i.d points from $\mathcal{D}$, we'll denote the set of samples S, $S \sim, \mathcal{D}$:

$$\widehat{\mu_{\mathrm{Rob}}(h)} = \frac{1}{|S|} \sum_{\substack{\mathbf{x}\in S\\ \mathbf{y}\sim N_\rho(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Individual Fairness: Likewise, individual fairness estimation given by random estimator is:

$$\widehat{\mu_{\mathrm{IFair}}(h)} = \frac{1}{|S|} \sum_{\substack{\mathbf{x}\in S \\ \mathbf{y}\sim N_{\rho,l}(\mathbf{x})}} \mathbb{1}_{h(\mathbf{x})\neq h(\mathbf{y})}$$

Group Fairness: Let S^+ denotes i.i.d samples from the first protected group and S^- i.i.d samples from the second protected group. Group Fairness (with demographic parity measure) is defined as:

$$\widehat{\mu_{\mathrm{GFair}}(h)} = \frac{1}{|S^+|} \sum_{\mathbf{x}\in S^+} \mathbb{1}_{h(\mathbf{x})=1} - \frac{1}{|S^-|} \sum_{\mathbf{x}\in S^-} \mathbb{1}_{h(\mathbf{x})=1}$$

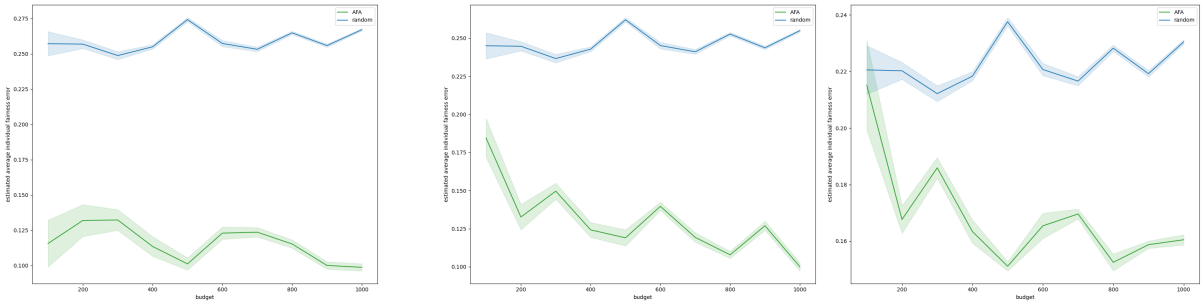
B. More experiments:

Individual fairness property depend on two perturbation parameters: ρ and l, we have seen in the experiments section that AFA is sensitive to high perturbation values that change the distribution on sample space ($\rho \approx -1$). Robustness and individual fairness make sense for reasonable values of ρ as discussed in the experiments section. However, the perturbation parameter l is a free parameter in the definition of individual fairness for which Hamming distance measures similarity between individuals. The parameter l answers the question: *What degree of similarity should the model refrain from distinguishing?* Hence, a good auditor would have the same performance for all possible values of the parameter l. To evaluate that, we fix $\rho = 0.30$ and we compare AFA and random estimator performances for a range of values of parameter l. Experiments details are summarized in table 2.

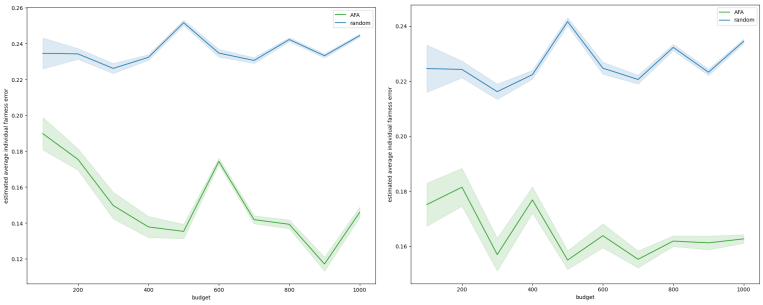
l-parameter	AFA μ_{IFair} error	random μ_{IFair} error
11	**0.123**	0.267
10	**0.119**	0.254
7	**0.141**	0.244
5	**0.169**	0.230
3	**0.166**	0.222

Table 2. A summary of theoretical results: This table summarizes the expression of the estimation for each property with query complexity and computational complexity.

As Figure 9 shows, AFA always outperform random estimator for the property of individual fairness for all different values of perturbation parameter l.



(a) Individual fairness error for $l = 11$ (b) Individual fairness error for $l = 10$ (c) Individual fairness error for $l = 7$



(d) Individual fairness error for $l = 5$ (e) Individual fairness error for $l = 3$

Figure 9. Comparison of individual fairness error over query budgets on COMPAS dataset.