Comparative performance evaluation across models. Table 1 summarizes the performance of AFA compared to baseline methods across various black-box models. We evaluated the performance using Logistic Regression that was trained using L_2 regularization, Multi-Layer Perceptron with one hidden layer and 100 neurons, and Random Forest model with 100 trees. AFA outperforms the baselines in each case.

Table 1: Average estimation error for statistical parity across different ML models. '—' denotes when a method cannot scale to the model.

Model	μCAL	CAL	Randomized μCAL	Uniform	Inefficient AFA	AFA
Logistic Regression	0.315	0.315	0.312	0.077	0.012	**0.006**
MLP	—	—	—	0.225	0.149	**0.147**
Random Forest	—	—	—	0.077	0.012	**0.006**

Performance evaluation on mix-valued data. To evaluate AFA's performance on continuous normalized data, we extended our experiment to the student performance dataset[1], where gender is the sensitive attribute. The student performance dataset consists of students' grades in various subjects. In this analysis, we fix gender as the protected attribute and audit the statistical parity for this attribute.

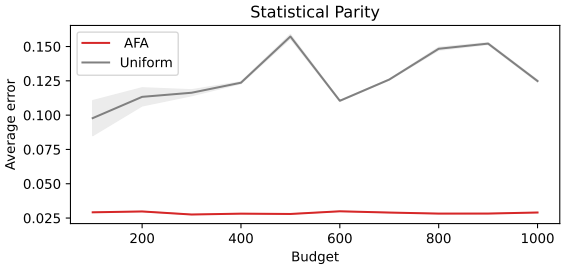

Figure 1: Error of AFA and uniform estimator auditors in estimating statistical parity.

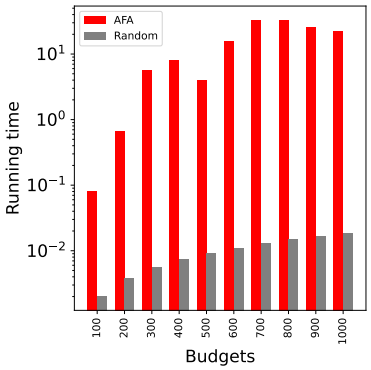

Figure 2: Running time for AFA and uniform estimator for statistical parity.

Figure 1 demonstrates that AFA outperforms the uniform estimator baseline. Figure 2 shows that our algorithm efficiently utilizes memory to compute Fourier weights, enabling the identification of significant Fourier coefficients.

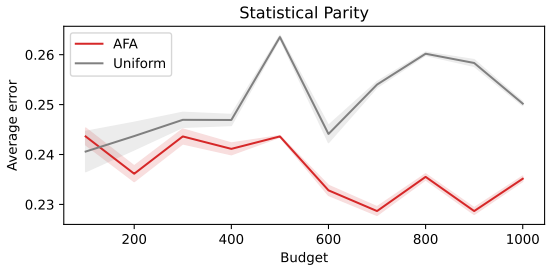

Figure 3: Error of AFA and uniform estimator auditors in estimating statistical parity for multi-label classifier.

Performance evaluation on multi-label classifiers. We further evaluate our algorithm using the Drug Consumption dataset [2], where the output is categorical and the protected attribute is gender. We train a logistic regression model on this dataset and audit its statistical parity in a black-box setting. Figure 3 demonstrates that our auditor outperforms the baseline in the multi-class setting

[1] https://www.kaggle.com/datasets/impapan/student-performance-data-set
[2] https://www.kaggle.com/datasets/obeykhadija/drug-consumptions-uci/data