A MORE ON FAIRCOCCO

A.1 CLOSED FORM EXPRESSION

We introduced covariance operators on RKHSs, which can be used to quantify unconditional $\left(V_{\hat{Y}A}\right)$ and conditional fairness $\left(V_{\hat{Y}\ddot{A}\mid Y}\right)$. `FairCOCCO` is based on the Hilbert-Schmidt (HS) norm of the covariance operators. An operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ is called HS if, for complete orthonormal systems $\{\phi_i\}$ of $\mathcal{H}_1$ and $\{\psi_j\}$ of $\mathcal{H}_2$, the sum $\sum_{i,j}\langle\psi_j, A\phi_i\rangle^2_{HS}$ is finite (Reed & Simon, 1980). Thus, for an HS operator A, the HS norm, $||A||_{HS}$ is defined as $||A||^2_{HS} = \sum_{i,j}\langle\psi_j, A\phi_i\rangle^2_{HS}$. Provided that $V_{\hat{Y}\ddot{A}\mid Y}$ and $V_{\hat{Y}A}$ are HS operators, `FairCOCCO` scores can be expressed as:

$$||V_{\hat{Y}\ddot{A}\mid Y}||^2_{HS} \quad \text{(conditional fairness measure)}$$
$$||V_{\hat{Y}A}||^2_{HS} \quad \text{(unconditional fairness measure)}$$

The umlaut on A represent extended variable sets, i.e. $\ddot{A} = (A, Y)$. Here, we briefly flesh out the closed-form expression of the empirical estimators, while more details can be found at (Fukumizu et al., 2007; Gretton et al., 2005). Let G_Y be the centered Gram matrices, such that:

$$G_{Y,ij} = \left\langle k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y^{(N)}, k_{\mathcal{Y}}(\cdot, Y_j) - \hat{m}_Y^{(N)} \right\rangle_{\mathcal{H}_{\mathcal{Y}}}$$

We choose a Gaussian RBF kernel, $k(Y_i, Y_j) = \exp\left(-\frac{||Y_i - Y_j||^2}{2\sigma^2}\right)\ \forall\, i, j \in N$, and employ the median heuristic introduced by Schölkopf et al. (2002), i.e. $\sigma = median\{|Y_i - Y_j|,\ \forall\, i \neq j \in N\}$ to select bandwidth σ. Additionally, $\hat{m}_Y^{(N)} = 1/N \sum_{i=1}^{N} k_{\mathcal{Y}}(\cdot, Y_i)$ is the empirical mean. $G_A, G_{\hat{Y}}$ are defined similarly. Based on this, proxy Gram matrices R_Y can be defined as follows:

$$R_Y = G_Y(G_Y + \epsilon N I_N)^{-1}$$

where $\epsilon = 1\text{e}{-4}$ is a regularization constant, used in the same way as Bach & Jordan (2002), I_N is an identity matrix and $R_{\hat{Y}}, R_A$ are defined similarly. The empirical estimator of $||\hat{V}_{\hat{Y}\ddot{A}\mid Y}||^2_{HS}$ can then be computed:

$$\hat{I} = ||\hat{V}_{\hat{Y}\ddot{A}\mid Y}||^2_{HS} \tag{12}$$
$$= \text{Tr}[R_{\hat{Y}}R_{\ddot{A}} - 2R_{\hat{Y}}R_{\ddot{A}}R_Y + R_{\hat{Y}}R_Y R_{\ddot{A}}R_Y] \tag{13}$$

The unconditional fairness score can similarly be estimated empirically as follows (note that unconditional dependence does not entail using extended variables):

$$\hat{I} = ||\hat{V}_{\hat{Y}A}||^2_{HS} \tag{14}$$
$$= \text{Tr}[R_{\hat{Y}}R_A] \tag{15}$$

Choice of Kernels. While, in general, kernel dependence measures depend not only on variable distributions, but also the choice of kernel, Fukumizu et al. (2007) showed that, in the limit of infinite data and assumptions on richness of the RKHS, the estimates converges to a kernel-independent value. We employ a Gaussian RBF (characteristic kernel) in our experiments.

On the computational complexity. For our experiments, we use a Gaussian RBF kernel: $k(X_i, X_j) = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right)\ \forall\, i, j \in N$ where σ is the tuneable bandwidth parameter. We employ the median heuristic introduced by Schölkopf et al. (2002), i.e. $\sigma = \text{median}\{|x_i - x_j|,\ \forall\, i \neq j \in N\}$ to select bandwidth.

As the calculation of (9) comprises a matrix inversion operation, the computational complexity scales with the number of samples in $\mathcal{O}(N^3)$. We improve the scaling with training samples in two ways, *(1)* by employing a low-rank Cholesky decomposition of the Gram matrix (of rank r), resulting in $\mathcal{O}(r^2N)$ complexity (Harbrecht et al., 2012) and *(2)* by estimating regulariser on mini-batches. We empirically demonstrate that these lead to strong results in real-world experiments.

A.2 FAIRCOCCO SCORE

Here, we derive `FairCOCCO` score from the underlying measure using the Cauchy-Schwarz Inequality. The `FairCOCCO` score for conditional fairness and unconditional fairness can be written as:

$$\texttt{FairCOCCO Score (unconditional)} = \frac{||V_{\hat{Y}A}||^2_{HS}}{||V_{\hat{Y}\hat{Y}}||_{HS}||V_{AA}||_{HS}}$$

$$\texttt{FairCOCCO Score (conditional)} = \frac{||V_{\hat{Y}\ddot{A}\,|\,Y}||^2_{HS}}{||R_{\hat{Y}} - R_{\hat{Y}}R_Y||_{HS}||R_{\ddot{A}} - R_{\ddot{A}}R_Y||_{HS}}$$

We start by looking unconditional version of `FairCOCCO`, we know from (14) and the Cauchy-Schwarz inequality for the inner-product $\langle \cdot, \cdot \rangle$ that:

$$\begin{aligned}
|||\hat{V}_{\hat{Y}A}||^2_{HS}| &= |\text{Tr}[R_{\hat{Y}}R_A]| = |\langle R^T_{\hat{Y}}, R_A\rangle| \\
&\leq ||R_{\hat{Y}}||_{HS}||R_A||_{HS} = \sqrt{\text{Tr}[R^T_{\hat{Y}}R_{\hat{Y}}]}\sqrt{\text{Tr}[R^T_A R_A]} \\
&= ||\hat{V}_{\hat{Y}\hat{Y}}||_{HS}||\hat{V}_{AA}||_{HS}
\end{aligned}$$

By the inequality, `FairCOCCO` Score (unconditional) $\in [-1, 1]$. Additionally, as the score is also non-negative, it takes value $\in [0, 1]$ where 0 indicates perfect fairness (as indicated by Lemma 3.1). By contrast, the score takes value 1 iff the gram matrices, $R_{\hat{Y}}$ and R_A, are linearly dependent (i.e. perfectly unfair). The derivation and interpretation can similarly be shown for the conditional case:

$$\begin{aligned}
|||\hat{V}_{\hat{Y}A\,|\,Y}||^2_{HS}| &= |\text{Tr}[R_{\hat{Y}}R_A - 2R_{\hat{Y}}R_AR_Y + R_{\hat{Y}}R_YR_AR_Y]| \\
&= |\text{Tr}[(R_{\hat{Y}} - R_{\hat{Y}}R_Y)(R_A - R_AR_Y)]| = |\langle (R_{\hat{Y}} - R_{\hat{Y}}R_Y)^T, (R_A - R_AR_Y)\rangle| \\
&\leq ||R_{\hat{Y}} - R_{\hat{Y}}R_Y||_{HS}||R_A - R_AR_Y||_{HS}
\end{aligned}$$

Here, $R_{\hat{Y}} - R_{\hat{Y}}R_Y$ is related to the conditional covariance operator, i.e. $\hat{V}_{\hat{Y}\hat{Y}\,|\,Y}$, which captures the conditional covariance of $\hat{Y}$ given Y. See (Fukumizu et al., 2007; 2009; Baker, 1973) and others for more.

B EXPERIMENTAL DETAILS

B.1 SUPERVISED LEARNING TASKS

B.1.1 MODEL DETAILS

For all experiments, we train a two-layer neural network with ReLU-activated nodes. The number of nodes chosen is between $40\sim100$ depending on the complexity of the data. The network is trained with Cross Entropy or MSE Loss and is optimized using Adam (Kingma & Ba, 2014). The hyperparameters include batch size $\in \{64, 128, 256\}$, learning rate $\in \{1\text{e}{-2}, 1\text{e}{-3}, 1\text{e}{-4}\}$, and fairness penalty $\in \{0.0, 0.5, 1.0, 2.0, 5.0\}$ and are chosen through cross-validation. For datasets without a defined test set, the data is split 60-20-20 into train, validation and test set and results are averaged over 10 runs. Experiments are run on either a CPU or NVIDIA Tesla K40C GPU, taking around an hour.

B.1.2 DATASETS

Adult (Kohavi, 1996). The task on the Adult dataset is to classify whether an individual's income exceeded \$50K/year based on census data. There are 48842 training instances and 14 attributes, 4 of which are sensitive attributes (`age`, `race`, `sex`, `native-country`). Here, the sensitive attribute is chosen to be `sex`, which can be either female or male.

Drug Consumption (Drugs) (Mirkes, 2015). The classification problem is whether an individual consumed drugs based on personality traits. The dataset contains 1885 respondents and 12 personality measurements. Respondents are questioned on drug use on 18 drugs, including a fictitious drug `Semeron` to identify over-claimers. Here, we focus on `Heroin` use, drop the respondents who

Table 7: **Description of datasets.** '-B' suffix indicates binary variables, '-D' indicates discrete variables (i.e. >2 classes) '-C' indicates continuous variables.

	Dataset	Examples	Features	Sensitive (A)	Outcome (Y)
Single sensitive attributes	**Adult**	45222	12	Gender-B	Income-B
	Drugs	1885	11	Ethnicity-B	Drug use-B
	German	1700	20	Foreign-B	Income-B
	COMPAS	6172	10	Ethnicity -B	Recidivism-B
Multiple sensitive attributes	**C&C**	1994	128	Ethnicity-C ($\times 4$)	Crime rate-C
	Students	649	33	Age-C, Gender-B	Performance-C
	KDD-Census	299285	40	Sex-B, Race-B, Age-C	Income-B
	Credit Card	30000	24	Sex-B, Marriage-D, Education-D	Default-B
	Law School	20798	12	Male-B, Race-D	Pass-B

claimed to use `Semeron` and transform the categorical response into a binary outcome: "Never Used" versus "Used". The binary sensitive attribute is `Ethnicity`.

South German Credit (German) (Hoffman, 1994). The German dataset contains 1000 instances with 20 predictor variables of a debtor's financial history and demographic information, which are used to predict binary credit risk (i.e. complied with credit contract or not). The sensitive attribute is a binary variable indicating whether the debtor is of foreign nationality.

COMPAS (Angwin et al., 2016). COMPAS is a commercial software commonly used by judges and parole officers for scoring a criminal defendant's likelihood of recidivism. The dataset contains 6172 instances with 10 features. The outcome is a binary variable corresponding to whether violent recidivism occurred (`is_violent_recid`) and the sensitive attribute is `race`, which is binarised into "Caucasian" and "Non-Caucasian" defendants.

Communities and Crime (C&C) (Redmond, 2009). C&C contains socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey and crime data from 1995 FBI UCR. It contains 1994 instances of communities with 128 attributes. The outcome of the regression problem is crime rate within each community `ViolentCrimesPerPop`, which is a continuous value. There are three sensitive attributes, corresponding to ethnic proportions in the community—`racePctBlack`, `racePctWhite`, `racePctAsian`.

Student Performance (Students) (Cortez, 2014). The Students dataset predicts academic performance in the last year of high school. There are 649 instances with 33 attributes, including past academic information and student demographics. The response variable is a continuous variable corresponding to final grade and the sensitive attributes are `age` (continuous value from 15-22) and `sex` ('F'-female, 'M'-male).

B.2 Time Series Task

The data used to develop and evaluate our experiment on fair imitation learning is extracted from the MIMIC-III ICU database (Johnson et al., 2016a), based on the Sepsis-3 cohort defined by Komorowski et al. (2018).

Discrimination in Healthcare. Sepsis is one of the leading causes of mortality in intensive care units (Singer et al., 2016), and while efforts have been made to provide clinical guidelines for treatment, physicians at the bedside largely rely on experience, giving rise to possible variations in fair treatments. Prejudice in healthcare has been reported in many instances—for example, healthcare professionals are more likely to downplay women's health concerns (Rogers & Ballantyne, 2008) and racial biases affect pain assessment and treatment prescribed (Hoffman et al., 2016). Thus, it is critical, when learning to imitate expert policy, that no underlying prejudices are leaked into the learned policy.

Problem Setup. We have access to a set of expert trajectories $\mathcal{D} = \{\tau_1, ..., \tau_N\}$, where each trajectory is a sequence of state-action pairs $\{(s_1, a_1), ..., (s_T, a_T)\}$. The time-varying state space is modelled with a Markov Decision Process (MDP), i.e. at every time step t, the agent observes current state s_t and takes action a_t.

Data. We obtain data from MIMIC-III and use the pre-processing scripts provided by Komorowski et al. (2018) to extract patients satisfying the Sepsis-3 criteria. For each patient, we have relevant physiological parameters, including demographics, lab values, vital signs and intake/output events. Data are aggregated into 4 hour windows.

State Space. The pre-processing yields 45×1 feature vectors for each patient at each time step, which are summarized in Table 8. We consider gender as the sensitive attribute.

Table 8: **MIMIC-III Features.** Description of patient features recorded at four hour intervals.

Feature Type	Features
Demographic	Gender, Age, Weight (kg),
Static	Re-admission, Glasgow Coma Scale (GCS), Sequential Organ Failure Assessment (SOFA), Systematic Inflammatory Response Syndrome (SIRS), Shock Index,
Lab Values	Potassium, Sodium, Chloride, Glucose, Magnesium, Calcium, White Blood Cell Count, Platelets Count, Bicarbonate, Hemoglobin, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), Arterial pH, Arterial Blood Gas, Arterial Lactate, Blood Urea Nitrogen (BUN), Creatinine, Serum Glutamic-Oxaloacetic Transaminase (SGOT), Serum Glutamic-Pyruvic Transaminase (SGPT), Total Bilirubin, International Normalized Ratio (INR),
Vitals	Heart Rate, Systolic Blood Pressure, Mean Blood Pressure, Diastolic Blood Pressure, Respiratory Rate, Temperature (Celsius), FiO2, PaO2, PaCO2, PaO2/FiO2 ratio, SpO2,
Intake/Output	Mechanical Ventilation, Fluid Intake (4 hourly), Fluid Intake (Total), Fluid Output (4 hourly), Fluid Output (Total)

Action Space. We define a binary action for medical intervention based on intravenous (IV) fluid and maximum vasopressor (VP) dosage in a given 4 hour window, where $a_t = 1$ represent either or both interventions taken, and $a_t = 0$ indicates no action taken.

Treatment Outcome. The ground truth treatment outcome in each time step is evaluated using SOFA (measuring organ failure) and the arterial lactate levels (higher in septic patients). Specifically, the treatment outcome penalizes high SOFA scores and increases in SOFA and lactate levels from the previous time step (Raghu et al., 2017):

$$Y_t = -0.025\mathbb{1}(s_{t+1}^{SOFA} = s_t^{SOFA} \;\&\; s_{t+1}^{SOFA} > 0) - 0.125(s_{t+1}^{SOFA} - s_T^{SOFA}) - 2\tanh(s_{t+1}^{lactate} - s_t^{lactate})$$

Behavioral Cloning. Our proposed framework should work with any imitation learning algorithm as long as predictions of action rewards are differentiable. For now, we will focus on behavioral cloning. The expert's demonstrations $\mathcal{D}$ are divided into i.i.d. state-action pairs. We train a neural network as described in the experimental setup to predict posterior action probabilities.

C Additional Experiments

In this section, we provide additional results to comprehensively evaluate our proposed methods, specifically:

1. **DP and EO**: While the main paper investigates fairness using EO, Appendix C.1 demonstrates application of `FairCOCCO` using DP and CAL notions of fairness.
2. **Estimation convergence**: Appendix C.2 evaluates the convergence of `FairCOCCO score` estimation on different mini-batch sizes on real datasets.
3. **Statistical testing**: Appendix C.3 demonstrates how the `FairCOCCO Score` can be employed as a test statistic in permutation-based testing for stronger fairness transparency.
4. **Sensitivity**: Appendix C.4 investigates performance sensitivities, specifically performance-fairness trade-offs, according to varying numbers of sensitive attributes.

C.1 Additional Results: Experiments with DP and CAL

To highlight `FairCOCCO`'s compatibility with fairness definitions other than EO, we apply it to demographic parity (DP) and calibration (CAL). We perform the same experiments on 1) binary classification tasks, 2) regression task with multiple sensitive attributes. The experiments are performed using the procedures described in the experimental setup.

Demographic Parity. DP requires statistical independence between predictions and attributes. *Disparate impact* (DI) is a metric frequently used to evaluate DP (Feldman et al., 2015):

$$DI = \frac{P(\hat{Y} = 1|A = 1)}{P(\hat{Y} = 1|A = 0)} \tag{16}$$

where $A = 1$ and $A = 0$ denote respectively the discriminated and non-discriminated groups. The US Equal Employment Opportunity Commission Recommendation advocates that DI should not be below 80%, commonly known as the 80%-rule.[4] DI closer to 1 corresponds to lower levels of disparate impacts across population subgroups. We show the performance of FairCOCCO for DP in Table 9 and 10, demonstrating superior performance on a benchmark of binary classification tasks as well as protection of multiple sensitive attributes in regression settings.

Table 9: **Performance in binary setting.** Accuracy (ACC) and DI under DP. *NN* is an unregularised neural network that is used as base learner; the best results are emboldened.

Method	COMPAS		German		Drug		Adult	
	ACC	DI	ACC	DI	ACC	DI	ACC	DI
Donini et al. (2018)	0.70 ± 0.02	0.81 ± 0.03	0.70 ± 0.06	0.93 ± 0.07	0.74 ± 0.03	0.75 ± 0.01	0.72	0.84
NN	0.90 ± 0.02	0.39 ± 0.32	0.74 ± 0.07	1.26 ± 0.54	0.80 ± 0.08	0.42 ± 0.22	0.84	0.22
Mary et al. (2019)	0.87 ± 0.04	0.76 ± 0.07	0.71 ± 0.08	0.96 ± 0.25	0.80 ± 0.06	0.73 ± 0.17	0.79	0.83
Steinberg et al. (2020b)	0.86 ± 0.03	0.83 ± 0.05	0.71 ± 0.06	0.93 ± 0.13	0.77 ± 0.03	0.86 ± 0.05	0.77	0.76
FairCOCCO	$\mathbf{0.88 \pm 0.03}$	$\mathbf{0.90 \pm 0.06}$	$\mathbf{0.73 \pm 0.06}$	$\mathbf{1.02 \pm 0.19}$	$\mathbf{0.78 \pm 0.02}$	$\mathbf{0.84 \pm 0.07}$	**0.83**	**0.97**

Table 10: **Protection of multiple attributes.** Level of protection provided to individual attributes when all attributes are simultaneously protected under DP. Lowest MSE & FairCOCCO scores are emboldened. **(left)** C&C dataset, **(right)** Students dataset.

Method	Joint		raceBlack	raceWhite	raceAsian	raceHisp
	MSE	COCCO	COCCO	COCCO	COCCO	COCCO
NN	0.22 ± 0.01	0.20 ± 0.08	0.16 ± 0.06	0.24 ± 0.03	0.03 ± 0.01	0.09 ± 0.05
FACL	0.53 ± 0.04	0.09 ± 0.02	0.07 ± 0.01	0.15 ± 0.04	0.05 ± 0.03	0.07 ± 0.02
FARMI	0.60 ± 0.07	0.12 ± 0.03	0.15 ± 0.02	0.15 ± 0.02	0.04 ± 0.01	0.06 ± 0.03
FairCOCCO	**0.49 ± 0.06**	**0.08 ± 0.02**	**0.05 ± 0.01**	**0.07 ± 0.02**	**0.03 ± 0.01**	**0.04 ± 0.01**

Method	Joint		age	sex
	MSE	COCCO	COCCO	COCCO
NN	0.25 ± 0.05	0.16 ± 0.06	0.13 ± 0.03	0.11 ± 0.07
FACL	0.30 ± 0.02	0.08 ± 0.01	0.04 ± 0.01	**0.03 ± 0.02**
FARMI	0.35 ± 0.05	0.11 ± 0.03	0.09 ± 0.02	0.05 ± 0.01
FairCOCCO	**0.33 ± 0.02**	**0.06 ± 0.02**	**0.03 ± 0.01**	0.04 ± 0.02

Calibration. CAL requires conditional independence between target and sensitive attributes given predictions. As the conditioning variable is continuous, we report the FairCOCCO score on the same experiments. We see in Table 11 and 12 that FairCOCCO achieves superior fair and predictive outcomes under different definitions of fairness when compared to other methods.

Table 11: **Performance in binary setting.** Accuracy (ACC) and FairCOCCO (COCCO) under CAL; the best results are emboldened.

Method	COMPAS		German		Drug		Adult	
	ACC	COCCO	ACC	COCCO	ACC	COCCO	ACC	COCCO
Donini et al. (2018)	0.76 ± 0.03	0.12 ± 0.02	0.70 ± 0.05	0.06 ± 0.01	0.80 ± 0.07	0.13 ± 0.21	0.78	0.16
NN	0.90 ± 0.02	0.07 ± 0.02	0.74 ± 0.07	0.07 ± 0.03	0.80 ± 0.08	0.24 ± 0.08	0.84	0.18
Mary et al. (2019)	0.87 ± 0.12	0.07 ± 0.03	0.71 ± 0.11	0.06 ± 0.02	$\mathbf{0.79 \pm 0.03}$	$\mathbf{0.08 \pm 0.03}$	0.81	0.15
(Steinberg et al., 2020b)	0.88 ± 0.03	0.06 ± 0.01	$\mathbf{0.73 \pm 0.06}$	0.04 ± 0.02	0.77 ± 0.05	0.16 ± 0.05	0.80	0.14
FairCOCCO	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.02 \pm 0.02}$	0.71 ± 0.05	$\mathbf{0.02 \pm 0.01}$	0.78 ± 0.06	0.11 ± 0.06	**0.83**	**0.11**

C.2 FAIRCOCCO ESTIMATION

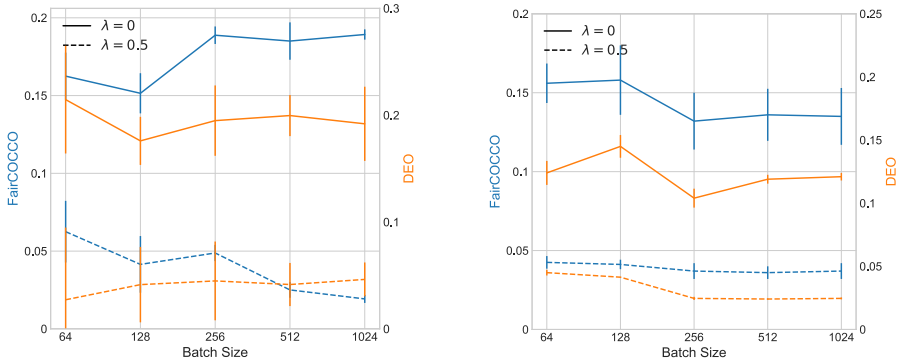
In this section, we provide additional results on convergence of FairCOCCO Score estimation as a function of batch size, similar to the experiment performed in the main paper. We show convergence on **Adult** and **German** dataset in Figure 2. We note that while convergence of estimation depends on properties of different datasets, the estimation of FairCOCCO Score stabilizes at batch sizes > 256.

[4] www.uniformguidelines.com.

Table 12: **Protection of multiple attributes**. Level of protection provided to individual attributes when all attributes are simultaneously protected under CAL. Lowest MSE and FairCOCCO score are emboldened. **(left)** C&C dataset, **(right)** Students dataset.

Method	Joint MSE	Joint COCCO	racePctBlack COCCO	racePctWhite COCCO	racePctAsian COCCO	racePctHisp COCCO
NN	0.22 ± 0.01	0.16 ± 0.03	0.16 ± 0.04	0.13 ± 0.03	0.07 ± 0.08	0.12 ± 0.03
FACL	0.55 ± 0.10	0.14 ± 0.02	0.11 ± 0.01	0.09 ± 0.03	0.11 ± 0.01	0.09 ± 0.04
FARMI	0.53 ± 0.05	0.15 ± 0.05	0.13 ± 0.02	0.12 ± 0.03	0.05 ± 0.01	0.10 ± 0.03
FairCOCCO	**0.47 ± 0.09**	**0.06 ± 0.01**	**0.08 ± 0.01**	**0.07 ± 0.02**	**0.03 ± 0.01**	**0.06 ± 0.01**

Method	Joint MSE	Joint COCCO	age COCCO	sex COCCO
NN	0.25 ± 0.05	0.11 ± 0.05	0.09 ± 0.01	0.05 ± 0.06
FACL	0.32 ± 0.03	0.14 ± 0.02	0.12 ± 0.02	0.07 ± 0.02
FARMI	0.36 ± 0.06	0.15 ± 0.01	0.11 ± 0.02	0.10 ± 0.02
FairCOCCO	**0.37 ± 0.05**	**0.04 ± 0.02**	**0.06 ± 0.01**	**0.03 ± 0.03**



(a) (b)

Figure 2: **Estimation of FairCOCCO Score. (a)** Adult dataset, **(b)** German dataset.

C.3 Statistical Testing

We demonstrate how the proposed fairness measures can be employed as a test statistic to perform statistical tests, resulting in stronger guarantees and transparency (Fukumizu et al., 2007; Gretton et al., 2005). We highlight that while other fairness measures (MI and MCC) can be developed as test statistics, the empirical estimation of these measures involve multiple levels of approximations, and it is unclear whether the approximated statistics still retain the theoretical properties. Figure 3 shows the distributions of predictions with fairness regularization. Notably, EO only requires statistical independence between predictions and sensitive attributes given true outcome, whereas DP enforces "strict" independence between predictions and attributes.

Table 13: **Statistical testing.** Accuracy-fairness trade-offs under different fairness notions and corresponding test of statistical significance. **(left)** EO setting, **(right)** DP setting.

λ	ACC	DEO	COCCO	p-value
0.0	78.33	0.66	0.21	0.00
0.2	76.67	0.39	0.14	0.14
0.5	70.36	0.07	0.03	0.45
1.0	67.78	0.03	0.02	0.74
2.0	60.57	0.00	0.01	0.90

λ	ACC	DI	COCCO	p-value
0.0	78.33	3.05	0.07	0.00
0.2	72.56	1.54	0.03	0.04
0.5	69.33	1.77	0.01	0.09
1.0	67.38	1.13	0.01	0.14
2.0	64.60	0.92	0.00	0.27

As the null distribution is not known (Fukumizu et al., 2007), permutation testing is performed. Table 13 reveals the accuracy-fairness trade-offs and p-values under different regulation strengths. The p-values indicate the probability of observing the test statistic under null hypothesis of (conditional) independence. As we expect, stronger fairness regularization leads to lower levels of unfairness as measured by DI and DEO, as well as stronger guarantees in statistical tests. For example, at $\lambda = 2.0$, we can say with 90% chance that predictions are conditionally independent of sensitive attributes (under EO) or 27% chance that predictions are independent of sensitive attributes (under DP).

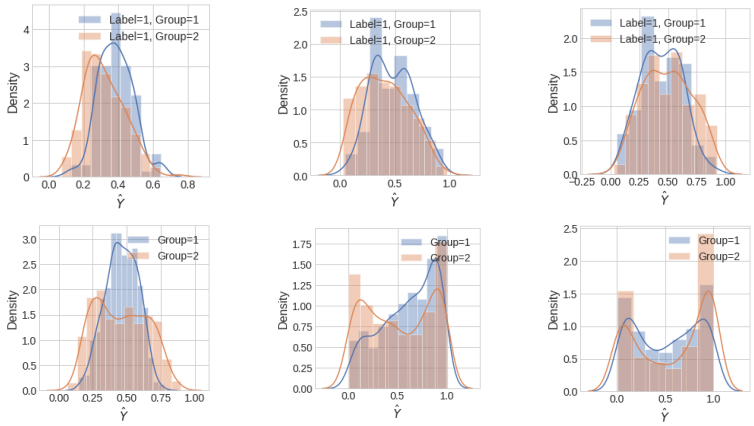


Figure 3: **Visualizing FairCOCCO regularization.** **(Top)** distribution of predictions for label 1 of different group memberships under EO. **(Bottom)** distribution of predictions for different group memberships under DP. Predictions are produced by regularized logistic regression model with $\lambda = 0, \lambda = 0.5, \lambda = 1.0$, respectively, across each row.

C.4 Sensitivity Analysis: Accuracy-Fairness Trade-offs

One of the key contributions of this study is the introduction of a differentiable fairness penalty that can naturally extend to multiple sensitive attributes. In this section, we generate the frontier of possible values on three experiments to better evaluate the sensitivity of our proposed methods to different numbers of sensitive attributes:

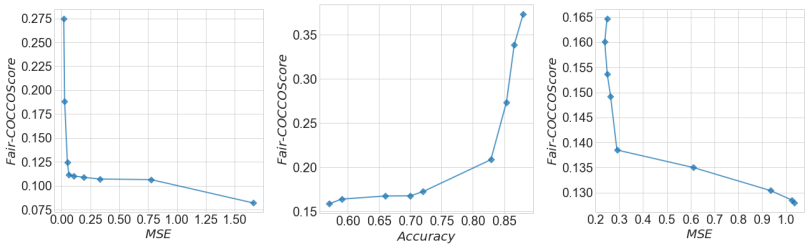


Figure 4: **Fairness-accuracy trade-off. (left)** C&C dataset with four sensitive attributes; **(middle)** students dataset with two sensitive attributes; **(right)** drugs dataset with three sensitive attributes.

- Regression on C&C with 4 attributes: `racePctBlack`, `racePctAsian`, `racePctWhite`, and `racePctHisp`,
- Regression on Students with 2 attributes: `age` and `gender`,
- Binary classification task on Drugs with 3 attributes: `age`, `gender`, and `ethnicity`.

As Figure 4 illustrates, similarly, fairness and prediction outcomes are achieved at various number of sensitive attributes.