Supplementary Material

The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective

Naman Goel,[1] **Alfonso Amayuelas,** [2] **Amit Deshpande,** [3] **Amit Sharma** [3]
[1] ETH Zürich, [2] EPFL Lausanne, [3] Microsoft Research
naman.goel@epfl.ch, amitdesh@microsoft.com, amshar@microsoft.com

A An example illustrating the causal graph framework for missing data

Consider the following example from (Mohan and Pearl 2020). We are interested in a dataset consisting of three variables - age (A), gender (G) and obesity (O). Figure 1a

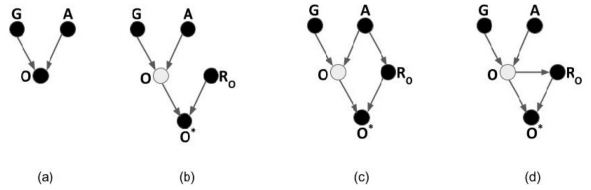


Figure 1: Causal graphs for missingness by Mohan and Pearl

shows the case of no missingness i.e. when all variables are fully observed. The edges between the variables shows the causal relationship between them. Figure 1b shows the case in which variable O has missingness and as a result we observe only O^*, where O^* = O if R_o = 0 and O^* = missing if R_o = 1. In this case of missingness, R_o is independent of other variables. In the language of missing data literature (Little and Rubin 2019), this is a case of *MCAR* (missing completely at random). Figure 1c shows the case when R_o is caused by A. This is a case of *MAR* (missing at random) because missingness is random conditioned on an observed variable A. Figure 1d shows the case when R_o is caused by variable O itself. This is a case of *MNAR* (missing not at random) because the variable O causes its own missingness.

B Proofs of Propositions 1-6

Before presenting the proofs, let us quickly revisit *d-separation* (Pearl 2009). d-separation answers the question whether (sets of) variables A and B are conditionally independent given C. If A and B are d-separated by C, then A and B are conditionally independent given C i.e. $A \perp\!\!\!\perp B|C$, otherwise independence is not guaranteed.

Definition 1. *Let A, B, C be the three non-intersecting subsets of nodes in a causal graph G. A path p is said to be d-separated (or blocked) by a set of nodes C if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in C, or*
2. *p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in C and such that no descendent of m is in C.*

A set C is said to d-separate A from B if and only if C blocks every path from a node in A to a node in B.

Proof of Proposition 1. To obtain the error rate estimate of a classifier, the standard procedure is to use the classifier to predict $\hat{Y}$ on i.i.d. samples of the data and compare it to Y. However, since the data is incomplete, we do not have access to $\hat{Y}$ and Y. Instead, we observe $\hat{Y^*}$ (predictions on the available samples) and we compare them to Y^* (outcomes for the available samples). Therefore, while the true error rate of the classifier for a group Z is $P(\hat{Y}|Y, Z)$, we end up estimating $P(\hat{Y}^*|Y^*, Z^*)$ due to incomplete data. We know that $P(\hat{Y}^*|Y^*, Z^*) = P(\hat{Y}|Y, Z, D = 1)$ (by definition).
It is easy to see that $P(\hat{Y}|Y, Z, D = 1) \neq P(\hat{Y}|Y, Z)$ because $\hat{Y}$ and D are not d-separated given Y, Z in Figure 1. There exists an active path $\hat{Y} \leftarrow X \to D$.

Proof of Proposition 2. In the causal graph shown in Figure 2a, we first show that Y and D are d-separated by X. There are two paths between Y and D - 1) $D \leftarrow X \to Y$ and $D \leftarrow Z \to X \to Y$. It is easy to see, based on Definition 1, that both of these paths are blocked by X. Thus, $Y \perp\!\!\!\perp D|X$. This means, $P(Y|X, D = 1) = P(Y|X)$. We know that, by definition, $P(Y|X, D = 1) = P(Y^*|X^*)$. Thus, the estimate for $P(Y|X)$ that we obtain from the incomplete data is consistent. We can similarly see that $Y \perp\!\!\!\perp D|X, Z$ is also true, implying that $P(Y|X, Z)$ estimate is also consistent.
On the other hand, there is a direct edge from the variable X to its missingness mechanism D in the causal graph shown in Figure 2a. Theorem 2 of Mohan and Pearl (2020) implies that joint distribution $P(X)$ is therefore not recoverable.

Proof of Proposition 3. The proof of Proposition 3 follows similarly to the proof of Proposition 2. The difference in causal graph shown in Figures 2b is that the path

$D \leftarrow Z \rightarrow Y$ is not d-separated by X. Thus, conditional independence is not guaranteed ($Y \not\perp\!\!\!\perp D|X$).
Similar to the previous case, D and X have a direct edge implying non-recoverability of the joint distribution $P(X)$.

Proof of Proposition 4. The graph shown in Figures 2c and 2d contain unobservable variables U. For reasoning about d-separation, this variable can be removed and replaced by a bi-directed edge (Pearl 2009; Mohan and Pearl 2020) between X and D (in Figure 2c) and between D and Y (in Figure 2d). It is easy to see now that all paths between Y and D are d-separated by X, Z in Figure 2c. This implies that our estimate of $P(Y|X, Z)$ from the incomplete data are consistent. On the other hand, D and Y now have a direct edge between them in Figure 2d. Using Theorem 3 of Mohan and Pearl (2020), we can claim that the conditional distribution $P(Y|X, Z)$ is not recoverable.
D and X have a direct edge implying non-recoverability of the joint distribution $P(X)$.

Proof of Proposition 5. The proof follows in the same way as that of Proposition 4. The main idea is to see that, even in the presence of the new variable D_a, X continues to block all paths between Y and D in Figure 2e and the direct edge between Y and D continues to exist in Figure 2f.

Proof of Proposition 6. In Figure 3a, X_2 is d-separated from D_1 by X_1. Thus, our estimate of $P(X_2|X_1)$ from the incomplete data is consistent. Similarly, in Figure 3b, Y is d-separated from D_2 by X_1, X_2. Thus, our estimate of $P(Y|X_1, X_2)$ from the incomplete data is consistent.
It remains to show that $P(Y|X_1)$ can also be recovered.

$$\begin{aligned} P(Y|X_1) &= \sum_{X_2} P(Y, X_2|X_1) \\ &= \sum_{X_2} P(Y|X_2, X_1) \cdot P(X_2|X_1) \end{aligned}$$

Thus, $P(Y|X_1)$ can be calculated by writing it in terms of $P(Y|X_2, X_1)$ and $P(X_2|X_1)$, both of which can be consistently estimated as shown earlier.

C Constraints in DF^2 Algorithm

In this section, we describe how one can write demographic parity and equal opportunity fairness constraints in the $\boldsymbol{DF^2}$ algorithm.

Demographic Parity. Demographic parity constraint at stage i is given by

$$P(\hat{Y}_i = 1|Z = a) = P(\hat{Y}_i = 1|Z = b)$$

This constraint can be replaced by an empirical estimate as follows:

$$\frac{\sum_{j=1}^{n_i} D_i[j] \cdot \mathbb{1}_{z_j=b}}{\sum_{j=1}^{n_i} \mathbb{1}_{z_j=b}} = \frac{\sum_{j=1}^{n_i} D_i[j] \cdot \mathbb{1}_{z_j=w}}{\sum_{j=1}^{n_i} \mathbb{1}_{z_j=w}}$$

Here b and w represent two values of the sensitive attribute Z (for e.g. black and white). $\mathbb{1}$ is the indicator function, which takes value 1 if the condition in the subscript is true, and 0 otherwise. D_i and n_i are defined in the main text of the paper.

Equality of Opportunity. Equal opportunity constraint at stage i is given by

$$P(\hat{Y}_i = 1|Y = 1, Z = a) = P(\hat{Y}_i = 1|Y = 1, Z = b)$$

This constraint can be replaced by an empirical estimate as follows:

$$\frac{\sum_{j=1}^{n_i} D_i[j] \cdot \mathbb{1}_{z_j=b} \cdot P_j(Y|X_1, \ldots, X_i)}{\sum_{j=1}^{n_i} \mathbb{1}_{z_j=b} \cdot P_j(Y|X_1, \ldots, X_i)} = \frac{\sum_{j=1}^{n_i} D_i[j] \cdot \mathbb{1}_{z_j=w} \cdot P_j(Y|X_1, \ldots, X_i)}{\sum_{j=1}^{n_i} \mathbb{1}_{z_j=w} \cdot P_j(Y|X_1, \ldots, X_i)}$$

Note that the constraints never use $P(X)$. The summation over the population may cause a misconception that we are indeed indirectly using wrong $P(X)$. But note that the summation is over the individuals who appear during the decision-making phase and not over the individuals in the training data (which has missingness). Therefore, training data is used only to estimate $P(Y|X_1, \ldots, X_i)$. Also note that the constraints are linear in $D_i[j]$.

D Steps of DF^2 Algorithm in a Two Stage Process

Algorithm $\boldsymbol{DF^2}$ Stage 1

Input Initial pool of individuals, their feature set X_1, and risk scores $P(Y|X_1)$
Output Individuals to be forwarded to Stage 2

1: Solve optimization problem 1 for $i = 1$ and obtain optimal decision provability vector D_1 for the input pool of individuals.
2: Select individuals based on the optimal decision vector D_1 determined in Step 1.

Algorithm $\boldsymbol{DF^2}$ Stage 2

Input Individuals forwarded by Stage 1, their feature set X_1, X_2, and risk scores $P(Y|X_1, X_2)$
Output Finally selected individuals

1: Solve optimization problem 1 for $i = 2$ and obtain optimal decision provability vector D_2 for the input individuals.
2: Select individuals based on the optimal decision vector D_2 determined in Step 1.

E Proof of Theorem 1

We compare the solution returned by the DF^2 algorithm with the optimal solution. Let us first formally understand the optimal solution in a 2-stage process. DF^2 solves an optimization problem at both stages, maximizing precision subject to budget and fairness constraints. However, the solution returned by the final stage is optimal only with respect to the input that it receives from the previous stage. On the other hand, the (absolute) optimal solution would be the one calculated by solving the same optimization problem, but with the entire input (i.e. without any filtering done by the first stage). In other words, suboptimality occurs in our multi-stage process algorithm because the first stage (with

limited information $P(Y|X_1)$) may filter out candidates that the final stage (with more information $P(Y|X_1, X_2)$) might have considered as better. Let α_1 and α_2 be the budget constraints in the two stages.

Let r_1 and r_2 denote the random variables $P(Y|X_1)$ and $P(Y|X_1, X_2)$. We know from (Menon and Williamson 2018; Corbett-Davies et al. 2017) that optimization problem of the form used in DF^2 results in a classifier where the first and the second stage classifiers apply subgroup specific thresholds on the risk scores r_1 and r_2, respectively (i.e. individuals above their subgroup-specific threshold are selected and everyone else is rejected); see Theorem 3.2 in (Corbett-Davies et al. 2017), whose proof works mutatis mutandis for our case with a budget constraint. We use the standard convention of calling $P(Y|X_1, \ldots X_i)$ as risk scores. It is just the probability of the outcome of interest given the observable features. In hiring example, this is the probability of an individual being good; everyone above a certain value for this probability may thus be shortlisted and everyone else be rejected. In recidivism example, this is the probability of an individual committing a crime again; everyone above a certain value for this probability may thus be detained and everyone else be released. In loan example, this is the probability of an individual being returning a loan; everyone above a certain value for this probability may thus be given loan and everyone else be denied. Let $\delta_z(\alpha_1)$ be the threshold (for a subgroup z) on r_1, and let $\delta_z(\alpha_2)$ be the threshold on r_2 assuming that the first stage doesn't exist (i.e. all individuals from the first stage are made available to the second stage).

Demographic Parity Constraints.

When the fairness constraints are demographic parity constraints, the optimization problem has to select candidates from both the groups with equal probability. This means if the budget constraint is α in a given stage, it must select the best α fraction of input candidates from both the groups. Thus, the probability that the DF^2 algorithm returns a suboptimal solution is nothing but the probability that a candidate who would have passed the threshold $\delta_z(\alpha_2)$ for their respective group doesn't pass the first stage threshold $\delta_z(\alpha_1)$. Remember that $\delta_z(\alpha_2)$ is the threshold on r_2 assuming that the first stage doesn't exist (i.e. all individuals from the first stage are made available to the second stage). We thus have,

$$P(D^* \neq D) = P(r_1 < \delta_z(\alpha_1) | r_2 \geq \delta_z(\alpha_2))$$

Using the Coherent Feature assumption, we can rewrite this as:

$$P(D^* \neq D) \leq P(r_1 < \delta_z(\alpha_1) | r_2 = \delta_z(\alpha_2))$$

Equality of Opportunity Constraints.

When the fairness constraints are equality of opportunity, it is no longer necessary to select best α fraction of input candidates from both the groups, if the budget constraint in a given stage is α. The solution may select unequal fraction of best candidates from the two groups as long as equal opportunity and the overall budget constraints are satisfied. The probability that the DF^2 algorithm returns a suboptimal solution (a solution with lower expected utility than the optimal solution) is equal to the probability that a candidate who would have passed the threshold $\delta_z(\alpha_2)$ for their respective group doesn't pass the first stage threshold $\delta_z(\alpha_1)$ and that the candidate can't be replaced by another candidate of equal or better utility without violating equal opportunity constraints. We will show that the probability of the second event (i.e. that the candidate can't be replaced by another candidate of better utility without violating equal opportunity constraints) is 1. Let us first consider the possibility of finding a better replacement for the candidate from the same group. Since we know that the optimization problem returns solution that are threshold based rules, it is clearly not possible to find a better replacement because all better candidates from the same group are already marked as selected in a threshold based decision rule. Let us now consider the possibility of finding a better replacement for the candidate from the other group. This is also impossible. If it was possible to select a better candidate from the other group without violating equal opportunity constraints, that candidate (for whom we are finding a replacement) wouldn't be selected in the optimal solution (by definition of optimal solution) in the first place. Thus, in the case of equal opportunity constraints too, the probability that the DF^2 algorithm returns a suboptimal solution (a solution with lower expected utility than the optimal solution) is equal to the probability that a candidate who would have passed the threshold $\delta_z(\alpha_2)$ for their respective group doesn't pass the first stage threshold $\delta_z(\alpha_1)$. Thus, we obtain the same bound as for demographic parity.

Thus, we obtain the following for both types of fairness constraints:

$$P(D^* \neq D) \leq P(r_1 < \delta_z(\alpha_1) | r_2 = \delta_z(\alpha_2))$$

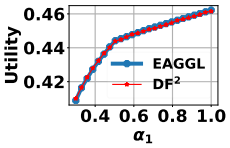
F Empirical Analysis - Additional Details

Dataset Details.

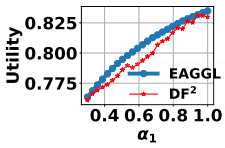
As mentioned in the paper earlier, we used same experimental conditions as Emelianov et al. (2019) by using their code available publicly on github *(https://github.com/vitaly-emelianov/ multistage fairness/)*. For completeness, we provide below the details of the datasets and pre-processing steps of Emelianov et al. (2019):

ADULT Dataset. The dataset (Dua and Graff 2017) contains 48842 rows and 14 features. The label *income* denotes if the salary is above $50,000$ dollars. In the experiments, Emelianov et al. (2019) binarize and use only the 6 following features: *sex* (is male), *age* (is above 35), *native-country* (from the EU or US), *education* (has Bachelor or Master degree), *hours-per-week* (works more than 35 hours per week) and *relationship* (is married).

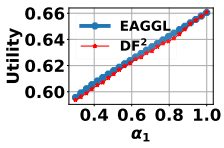
COMPAS Dataset. The COMPAS dataset (Larson et al. 2016) contains information about defendants, such as their name, gender, age, race, start of the sentence, end of the sentence, charge description etc. and a label *recidivism* denot-



(a) ADULT

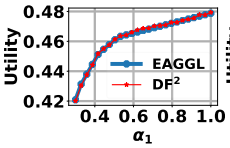


(b) GERMAN

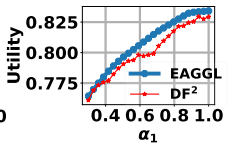


(c) COMPAS

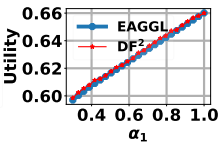
Figure 2: Demographic Parity Fairness Constraints (3-Stage Process)



(a) ADULT



(b) GERMAN



(c) COMPAS

Figure 3: Equality of Opportunity Fairness Constraints (3-Stage Process)

ing whether a defendant recidivates within 2 years or not. Emelianov et al. (2019) use only the rows for Caucasian and African-American defendants (total 6150 rows). They binarize and leave only the following 6 features: *sex* (is male), *young* (younger than 25), *old* (older than 45), *long sentence* (sentence was longer than 30 days), *drugs* (the arrest was due to selling or possessing drugs), *race* (is Caucasian).

GERMAN Dataset. The German Credit data from (Dua and Graff 2017) contains information about 1000 applicants for credit. The label *returns* shows if applicant payed back his loan. Emelianov et al. (2019) binarize and use 6 features: *job* (is employed), *housing* (owns house), *sex* (is male), *savings* (greater than 500 DM), *credit history* (all credits payed back duly), *age* (older than 50).

3-Stage Decision-making Process.

In addition to the 2-stage process discussed in the paper, we also simulated a 3-stage decision making process by setting $\alpha_3 = 0.3$, $\alpha_2 = 0.4$ and varying α_1 between 0.4 and 1. The sequence of feature observation in the three stages was same as (Emelianov et al. 2019) and is provided here for completeness. In the ADULT dataset, the first stage observes *sex* and *age*, the second stage adds *education* and *native country* to the previous features, and finally, the third stage adds *relationship*. In the COMPAS dataset, the first stage observes *race* and *young*, the second stage adds *drugs* and *old* to the previous features, and finally, the third stage adds *sex* and *long sentence*. In the GERMAN dataset, the first stage observes *sex* and *job*, the second stage adds *housing* and *credit history* to the previous features, and finally the third stage adds *age* and *savings*.

We make similar observations as in the 2-stage process. The results are shown in Figure 2 (for demographic parity constraints) and Figure 3 (for equal opportunity constraints). DF^2 algorithm obtains almost the same utility as the optimal *EAGGL* algorithm. In case of GERMAN dataset, we observe that the performance drops marginally. This may be due to the specific sequence of features observed in different stages in the GERMAN dataset (see Theorem 1) rather than the number of stages.

References

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.

Dua, D.; and Graff, C. 2017. UCI machine learning repository .

Emelianov, V.; Arvanitakis, G.; Gast, N.; Gummadi, K.; and Loiseau, P. 2019. The price of local fairness in multistage selection. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5836–5842. AAAI Press.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. https://github.com/propublica/compas-analysis, 2016. .

Little, R. J.; and Rubin, D. B. 2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 107–118.

Mohan, K.; and Pearl, J. 2020. Graphical models for processing missing data. *Journal of American Statistical Association(JASA)* .

Pearl, J. 2009. *Causality*. Cambridge university press.