A COUNTEREXAMPLE

In this section we show that in the multivariate setting, the worst-case counterfactual unfairness with a confounding budget of $p_{\text{max}}$ is not necessarily obtained when all non-zero entries of the correlation matrix are set to $p_{\text{max}}$. To this end, it suffices to find a symmetric matrix $A$ with 1s on the diagonal that is not positive-semidefinite when all its non-zero off-diagonal entries are set to the same value, which we define to be the considered confounding budget $p_{\text{max}}$. Since each valid correlation matrix must be positive-semidefinite, the correlation matrix for the worst-case counterfactual unfairness must be different from $A$ (while maintaining the zero entries). Because all off-diagonal entries are upper bounded by $p_{\text{max}}$, at least one of them must be smaller than the corresponding value in $A$.

For example, consider

$$A = \begin{pmatrix} 1 & p_{\text{max}} & p_{\text{max}} \\ p_{\text{max}} & 1 & 0 \\ p_{\text{max}} & 0 & 1 \end{pmatrix}.$$ 

Since the eigenvalues of $A$ are $1$, $1 - \sqrt{2p_{\text{max}}}$, and $1 + \sqrt{2p_{\text{max}}}$, we see that $A$ is not positive-semidefinite for $p_{\text{max}} > 1/\sqrt{2}$.

In general, the matrix $A \in \mathbb{R}^{n \times n}$ with $A_{ii} = 1$ for $i \in \{1, \ldots, n\}$, $A_{i1} = A_{1i} = p_{\text{max}}$ for $i \in \{2, \ldots, n\}$ and $A_{ij} = 0$ for all remaining entries, has the eigenvalues (without multiplicity) $1$, $1 - \sqrt{n-1}p$, and $1 + \sqrt{n-1}p$. Therefore, $A$ is not positive-semidefinite for $p_{\text{max}} > 1/\sqrt{n-1}$. We conclude that as the dimensionality of the problem increases, we may encounter such situations for ever smaller confounding budget.

B COMPUTATIONAL CONSIDERATIONS

Step 17 of Algorithm 1 is the main place where code optimization can take place, and alternatives to the (local) penalized maximum likelihood taking place there could be suggested (perhaps using spectral methods). It is hard though to say much in general about Step 20, as counterfactual fairness allows for a large variety of loss functions usable in supervised learning. In the case of linear predictors, it is still a non-convex problem due to the complex structure of the correlation matrix, and for now we leave as an open problem whether non-gradient based optimization may find better local minima.

C PATH-SPECIFIC SENSITIVITY

Path-specific effects were not originally described by Kusner et al. (2017) as the goal there was to introduce...