A COUNTEREXAMPLE

In this section we show that in the multivariate setting, the worst-case counterfactual unfairness with a confounding budget of $p_{\max}$ is not necessarily obtained when all non-zero entries of the correlation matrix are set to $p_{\max}$. 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 $p_{\max}$. 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 $p_{\max}$, at least one of them must be smaller than the corresponding value in A.

For example, consider

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

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

In general, the matrix $A \in \mathbb{R}^{n \times n}$ with $A_{ii} = 1$ for $i \in \{1, \ldots, n\}$, $A_{1i} = A_{i1} = p_{\max}$ 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 $p_{\max} > 1/\sqrt{n-1}$. 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 the core idea of counterfactual fairness in a way as accessible as possible (some discussion is provided in the supplementary material of that paper). See Chiappa & Gillam (2018) for one take on the problem, and Loftus et al. (2018) for another take to be fully developed in a future paper. Here we consider an example that illustrates how notions of path-specific effects (Shpitser, 2013) can be easily pipelined with our sensitivity analysis framework.

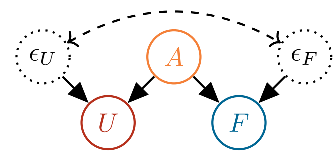

Figure 5: A path-specific model where the path from protected attribute A to feature U is unfair and the path from A to feature F is fair.

Consider Figure 5, where the path from $A \to U$ is considered unfair and $A \to F$ is considered fair, in the sense that we do not want a non-zero path-specific effect of A on $\hat{Y}$ that is comprised by a possible path $A \to U \to \hat{Y}$ in the causal graph implied by the chosen construction of $\hat{Y}$. Then a path-specific counterfactually fair predictor is one that uses $\{\epsilon_U, F\}$ as input. Note that the only difference this makes in our grid-based tool is that we only estimate the error ϵ_U for the unfair path in Model A (step 2, Section 3.3) and fit a predictor on $\{\epsilon_U, F\}$ (step 3, Section 3.3). Additionally, we only compute the incorrect error terms of the counterfactuals in Model B, using the weights of Model A (step 4, Section 3.4). For the optimization-based tool we would change lines 13, 14, and 20 in the same way.