# Supplemental Appendix for "Learning Optimal Fair Policies"

**Razieh Nabi** [1] **Daniel Malinsky** [1] **Ilya Shpitser** [1]

## Appendix A: G-estimation

G-estimation applies to structural nested models, which directly model the counterfactual deviations in outcome from a reference treatment value (which we take to be $A = 0$) conditional on history, assuming all future decisions are already optimal. Specifically, for each decision point $k$ we posit a *structural nested mean model (SNMM)* parameterized by $\psi$ as follows:

$$\gamma_k(H_k,a_k;\psi) = \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k],$$

where $\underline{A}_{k+1}$ represents all treatments administered from time $k+1$ onwards. In words, $\gamma_k$ is the contrast of the counterfactual mean (conditional on observed history $H_k$) where the past decisions are set to their observed values, the present decision is either $a_k$ or a reference decision $a_k = 0$, and all future decisions are made optimally, $f^*_{\underline{A}_{k+1}}$.

Note that if the true $\gamma_k(H_k, a_k; \psi)$ were known, the optimal treatment policies are those that maximize this "blip" function at each stage: $f^*_{A_k} = \arg\max_{a_k} \gamma_k(H_k, a_k; \psi)$. In order to estimate $\psi$ using data, let

$$U(\psi, \zeta(\psi), \alpha) = \sum_{k=1}^{K} \{G_k(\psi) - \mathbb{E}\left[G_k(\psi) \mid H_k; \zeta\right]\} \times \{d_k(H_k, A_k) - E\left[d_k(H_k, A_k) \mid H_k; \alpha\right]\}, \quad (1)$$

where $d_k(H_k, A_k)$ is any function of $H_k$ and $A_k$ and $G_k(\psi)$ is defined as

$$Y - \gamma_k(H_k, a_k; \psi) + \sum_{i=k+1}^{K} \left[\gamma_i(H_i, a_i^*; \psi) - \gamma_i(H_i, a_i; \psi)\right],$$

($a_i^*$ is the optimal decision at $i$th stage). Consistent estimators of $\psi$ can be obtained solving the estimating equations $\mathbb{E}[U(\psi, \zeta(\psi), \alpha)] = 0$, as shown in Robins (2004).

Both of the modifications discussed for Q-learning and value search must be applied when learning fair optimal policies by g-estimation. Specifically, we determine optimal polities not from the SNMM contrast $\gamma_k(H_k, a_k; \psi) = \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k]$ itself, but rather from a modified contrast $\gamma^*_k(H_k \setminus M, a_k; \psi) = \sum_{m,s} \gamma_k(H_k, a_k; \psi) p^*(M|S, X) p^*(S|X) = \mathbb{E}[Y(\bar{a}_{k-1}, a_k, f^*_{\underline{A}_{k+1}}) - Y(\bar{a}_{k-1}, a_k = 0, f^*_{\underline{A}_{k+1}}) \mid H_k \setminus \{M, S\}]$ which does not use $M$ and $S$. This is analogous to removing $M$ and $S$ from the Q-functions defined in Section 4 and is done for the same reason: $M, S$ are drawn from $p(Z)$, not $p^*(Z)$.

Second, the estimating equations for $\psi$ must use constrained models (in particular for $M$ and $S$), and must be empirically solved using observations only from $p^*(Z)$. As was done with value search, we solve equation (1) empirically using a dataset where each row $x_n, s_n, m_n$ is replaced by $I$ rows of the form $x_n, s^*_{ni}, m^*_{ni}$, $i = 1, \ldots, I$, with $s^*_{ni}$ and $m^*_{ni}$ drawn from $p^*(S|x_n; \alpha_s)$ and $p^*(M|x_n, S; \alpha_m)$, respectively.

## Appendix B: Simulation details and additional results on synthetic data

Here we report the precise parameter settings used in our simulation studies. The following regression models were used in our simulation study of the two-stage decision problem:

$$\begin{aligned}
X_1 &\sim |\mathcal{N}(0,1)| \\
(X_2, X_3) &\sim \mathcal{N}(0, \text{diag}(2)) \\
S &\sim \text{Bernoulli}(p = 0.5) \\
\text{logit}(p(M=1)) &\sim -1 + X_1 + X_2 + X_3 + S \\
&\quad + 3SX_1 + SX_2 + SX_3 \\
\text{logit}(p(A_1=1)) &\sim 1 - X1 + X2 + S + M - SX_1 + SX_2 \\
&\quad + MS - 3MX_1 + 0.5MX_2 \\
\text{logit}(p(Y_1=1)) &\sim -2 + X_1 + X_2 + S + M + A \\
&\quad + SX_2 + MS + AS + AM \\
\text{logit}(p(A_2=1)) &\sim 1 - X_1 + X_2 + M + A + W \\
&\quad + S(1 - X_1 + X_2 + M - A) \\
&\quad - 3MX_1 + 0.5MX_2 - AX_1 - AX_2 \\
Y &= 2.5 + X_1 + X_2 + M + W + B \\
&\quad + S(1 + X_1 + X_2 + M + A + W) \\
&\quad + A(1 + M - 2W) + MW \\
&\quad + B(-X_1 + 2X_2 - M) + WX_1 + \mathcal{N}(0,1)
\end{aligned}$$

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. Correspondence to: Razieh Nabi <rnabi@jhu.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

| | Unfair Policy | Fair Policy |
|---|---|---|
| **Q-learning** | $1.414 \pm 0.0056$ | $1.189 \pm 0.0059$ |
| **value search** | $1.134 \pm 0.0245$ | $1.056 \pm 0.0299$ |
| **g-estimation** | $1.375 \pm 0.0099$ | $1.312 \pm 0.0102$ |

*Table 1.* Comparison of population outcomes $\mathbb{E}[Y]$ under policies learned by different methods. The value under the observed policy was $0.24 \pm 0.006$.

For this two-stage setting we estimated the optimal policies using Q-learning and value search. In value search, we considered restricted class of polices of the form $p(A_1 = 1|X, S, M) = -1 + \alpha_x X + \alpha_s S + \alpha_m M + \alpha_{sx} SX + \alpha_{sm} SM + \alpha_{mx} MX$, and $p(A_2 = 1|X, S, M, A_1, Y_1) = -1 + \alpha_x X + \alpha_s S + \alpha_m M + \alpha_a A + \alpha_{y_1} Y_1 + \alpha_{sx} SX + \alpha_{sm} SM + \alpha_{mx} MX + \alpha_{as} AS + \alpha_{ax} AX$ where all $\alpha$s range from $-3$ to 3 by $0.5$ increments and estimated the value of policies for each combination of $\alpha$s using equation (7).

A third method for estimating policies is to directly model the counterfactual contrasts known as *optimal blip-to-zero functions* and then learn these functions by g-estimation (Robins, 2004); see Appendix A. We implemented our modified fair g-estimation for a single-stage decision problem and compared the results with Q-learning and value search. The results are provided in Table 1. The data generating process for the single-stage decision problem matches the causal model shown in Fig. 1(a) where $X$, $S$, $M$, and $A$ were generated the same way as described above. The outcome $Y$ was generated from a standard normal distribution with mean $-2 + X + S + M + A - 3SX_2 + MS + AS + AM + AX_2 + AX_3$. We used estimators in Theorem 1 to compute $\text{PSE}^{sy}$ and $\text{PSE}^{sa}$ which require using $M$ and $S$ models. In this synthetic data, the $\text{PSE}^{sy}$ was $1.618$ (on the mean scale) and was restricted to lie between $-0.1$ and $0.1$. The $\text{PSE}^{sa}$ was $0.685$ (on the odds ratio scale) and was restricted to lie between $0.95$ and $1.05$.

## Appendix C: Details and additional results on the COMPAS data experiment

The regression models we used in the COMPAS data analysis were specified as follows:

$$\begin{aligned}
\text{logit}(p(M = 1)) &\sim X_1 + X_2 + S + SX_1 + SX_2 \\
\text{logit}(p(A = 1)) &\sim X_1 + X_2 + S + M + SX_1 \\
&\quad + SX_2 + MS + MX_1 + MX_2 \\
Y &\sim X_1 + X_2 + S + M + A + SX_1 + SX_2 \\
&\quad + AS + AM + MS + MX_1 + MX_2 \\
&\quad + AX_1 + AX_2
\end{aligned}$$

For estimating the PSEs which we constrain, we used the same IPW estimators described in the main paper and reproduced in the theorem below. We constrained the PSEs to lie between $-0.05$ and $0.05$ and $0.95$ and $1.05$, respectively.

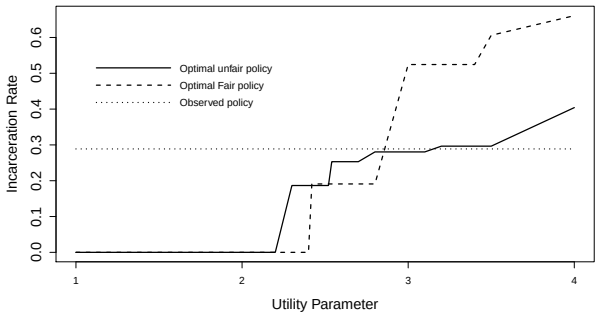


*Figure 3.* Overall incarceration rates for the COMPAS data as a function of the utility parameter $\theta$.

In Fig. 3, we compare the overall incarceration rates recommended by the optimal fair and unconstrained policies on the COMPAS data, as a function of the utility parameter $\theta$. For low values of $\theta$ the incarceration rate is zero, and becomes higher as $\theta$ increases, but differentially for the fair and unconstrained optimal policies. The difference between the policies depends crucially on the utility function. For some values of the utility parameter, the unfair and fair policies coincide, but for other values we would expect significantly different overall incarceration rates as well as different disparities between racial groups (see result in the main paper).

In Fig. 4, we show the relative utility achieved by the optimal fair and unconstrained policies, as well as the utility of the observed decision pattern, as a function of $\theta$. As expected, choosing an optimal policy improves on the observed policy, with the unfair (unconstrained) choice being higher utility than the fair (constrained) choice; we sacrifice some optimality to satisfy the fairness constraints. However, the difference depends on the utility parameter and for a range of parameter values the fair and unfair policies are nearly the same in terms of optimality (even when they may disagree on the resulting incarceration rate, around $\theta = 2.6$). The fair and unfair policies drift far apart in terms of utility around $\theta = 3$, when the policies recommend an incarceration rate comparable to or higher than the observed rate.

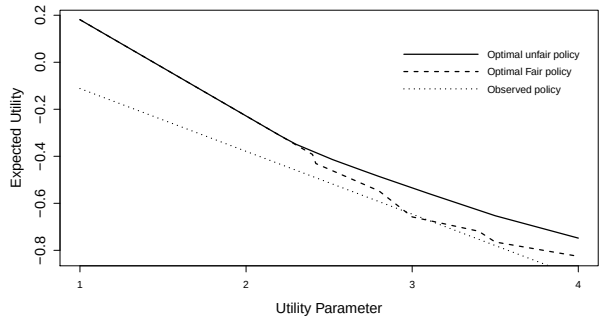


*Figure 4.* The relative utility of policies for the COMPAS data as a function of the utility parameter $\theta$.

## Appendix D: Proofs

**Theorem 1** *Assume $S$ is binary. Under the causal model above, the following are consistent estimators of PSE$^{sy}$ and PSE$^{sa_k}$, assuming all models are correctly specified:*

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n=s)}{p(S_n|X_n)}\frac{p(M_n|s',X_n)}{p(M_n|s,X_n)} - \frac{\mathbb{I}(S_n=s')}{p(S_n|X_n)}\right\}Y_n \quad (2)$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n=s)}{p(S_n|X_n)}\frac{p(M_n|s',X_n)}{p(M_n|s,X_n)} - \frac{\mathbb{I}(S_n=s')}{p(S_n|X_n)}\right\}A_{kn} \quad (3)$$

*Proof:* The latent projection (Verma and Pearl, 1990) of any $K$ stage DAG onto $X, S, M, A, Y$ suffices to identify and estimate the two path-specific effects in question, and this latent projection is the complete DAG with topological ordering $X, S, M, A, Y$. The consistency of the estimators above then follows directly from derivations in (Tchetgen Tchetgen and Shpitser, 2014). As an example, we have the following derivation for the first term of $g^{sy}(Z)$:

$$\begin{aligned}
&\sum_{X,M}\mathbb{E}\left[Y|s,M,X\right]p(M|s',X)p(X)\\
&=\sum_{X,M,A,Y} Y p(Y|s,M,A,X)p(A|s,M,X)p(M|s',X)p(X)\\
&=\sum_{X,S,M,A,Y}\frac{\mathbb{I}(S=s)p(M|s',X)}{p(S|X)p(M|s,X)}Y dp(Y,S,M,A,X)\\
&=\mathbb{E}\left[\frac{\mathbb{I}(S=s)p(M|s',X)}{p(S|X)p(M|s,X)}Y\right]
\end{aligned}$$

which is precisely the identifying functional for the first term of the PSE we are interested in. That the above estimator is consistent for this functional is a standard result. □

**Theorem 2** *Consider the K-stage decision problem described by the DAG in Fig 1c. Let $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$ be the constrained models chosen to satisfy PSE$^{sy}=0$ and PSE$^{sa_k}=0$. Let $\tilde{p}(Z)$ be the joint distribution induced by $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$, and where all other distributions in the factorization are unrestricted. That is,*

$$\tilde{p}(Z) \equiv p(X)p^*(S|X;\alpha_s)p^*(M|S,X;\alpha_m) \times \prod_{k=1}^{K} p(A_k|H_k)p(Y_k|A_k,H_k).$$

*Then the functionals PSE$^{sy}$ and PSE$^{sa_k}$ taken w.r.t. $\tilde{p}(Z)$ are also zero.*

*Proof:* Let $Y \equiv Y_K$. Because $M$ preceeds all $A_k, Y_k$ for $k = 1,\ldots K$, it suffices to consider the latent projection with only variables $X, S, M, A, Y$ without affecting identifiability considerations. Then we have the following:

$$\begin{aligned}
&\widetilde{\text{PSE}}^{sy}\\
&=\tilde{\mathbb{E}}[Y(s,M(s'))] - \tilde{\mathbb{E}}[Y(s')]\\
&=\sum_{X,M}\{\tilde{\mathbb{E}}[Y|s,M,X] - \tilde{\mathbb{E}}[Y|s',M,X]\}p^*(M|s',X;\alpha_m)p(X)\\
&=\sum_{X,M}\{\mathbb{E}[Y|s,M,X] - \mathbb{E}[Y|s',M,X]\}p^*(M|s',X;\alpha_m)p(X)\\
&=\sum_{X,M,Y} Y\{p(Y|s,M,X) - p(Y|s',M,X)\}p^*(M|s',X;\alpha_m)p(X)\\
&=\sum_{X,S,M,Y} Y\left\{\frac{\mathbb{I}(S=s)}{p^*(S|X;\alpha_s)}\frac{p^*(M|s',X;\alpha_m)}{p^*(M|s,X;\alpha_m)} - \frac{\mathbb{I}(S=s')}{p^*(S|X;\alpha_s)}\right\}\\
&\times p(Y|M,S,X)p^*(M|S,X;\alpha_m)p^*(S|X;\alpha_s)p(X)\\
&=0
\end{aligned}$$

by choice of $p^*(M|S,X;\alpha_m)$ and $p^*(S|X;\alpha_s)$. The proof is structurally the same for $\widetilde{\text{PSE}}^{sa_k}$. □

## Appendix E: Modified results with multiple sets of mediators

In the main paper, we discussed a $K$-stage decision problem with one set of permissible mediators, $M$. Here, we extend those results to the setting where we have multiple sets of mediators $M_1,\ldots,M_K$, i.e., a DAG with topological ordering $X, S, M_1, A_1, Y_1, \ldots, M_K, A_K, Y_K$. In this case, we consider the following paths impermissible: PSE$^{sy}$, representing the effect of $S$ on $Y$ along all paths *other than* the paths of the form $S \to M_k \to \ldots \to Y$ $(\forall k)$; and PSE$^{sa_k}$, representing the effect of $S$ on $A_k$ along all paths *other than* the paths of the form $S \to M_j \to \ldots \to A_k$ $(\forall j \leq k)$. That is, we consider *only* pathways connecting $S$ and $A_k$ or $Y$ through the allowed mediators $M_1,\ldots,M_K$ to be fair. In this case, the PSEs are identified by a modification of the

previous formula given in Section 3.2.

$$\begin{aligned}
&\text{PSE}^{sy} = \mathbb{E}[Y(s, M_1(s'), \ldots, M_K(s'))] - \mathbb{E}[Y(s')] \\
&= \sum_{x, \overline{m}_K, \overline{a}_{K-1}, \overline{y}_{K-1},} \{\mathbb{E}[Y|s, \overline{M}_K, \overline{A}_{K-1}, \overline{Y}_{K-1}, X] \\
&- \mathbb{E}[Y|s', \overline{M}_K, \overline{A}_{K-1}, \overline{Y}_{K-1}, X]\} \prod_{k=1}^{K} p(M_k|s', \overline{A}_{k-1}, \overline{Y}_{k-1}, X) \\
&\times \prod_{k=1}^{K-1} p(A_k|s, \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_k, X) p(Y_k|s, \overline{M}_k, \overline{A}_k, \overline{Y}_{k-1}, X) p(X)
\end{aligned}$$

$$\begin{aligned}
&\text{PSE}^{sa_k} = \mathbb{E}[A_k(s, M_1(s'), \ldots, M_K(s'))] - \mathbb{E}[A_k(s')] \\
&= \sum_{x, \overline{m}_k, \overline{a}_{k-1}, \overline{y}_{k-1},} \{\mathbb{E}[A_k|s, \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_{k-1}, X] \\
&- \mathbb{E}[A_k|s', \overline{M}_k, \overline{A}_{k-1}, \overline{Y}_{k-1}, X]\} \prod_{k=1}^{K} p(M_k|s', \overline{A}_{k-1}, \overline{Y}_{k-1}, X) \\
&\times \prod_{j=1}^{k-1} p(A_j|s, \overline{M}_j, \overline{A}_{j-1}, \overline{Y}_j, X) p(Y_j|s, \overline{M}_j, \overline{A}_j, \overline{Y}_{j-1}, X) p(X)
\end{aligned}$$

With these definitions, we can replace the estimators in Theorem 1 with:

$$\begin{aligned}
&\widehat{g}^{sy}(Z) = \\
&\frac{1}{N} \sum_{n=1}^{N} \Big\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \prod_{k=1}^{K} \frac{p(M_{k,n}|s', \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)}{p(M_{k,n}|s, \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)} \\
&- \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \Big\} Y_n
\end{aligned}$$

$$\begin{aligned}
&\widehat{g}^{sa_k}(Z) = \\
&\frac{1}{N} \sum_{n=1}^{N} \Big\{ \frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)} \prod_{k=1}^{K} \frac{p(M_{k,n}|s', \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)}{p(M_{k,n}|s, \overline{A}_{k-1,n}, \overline{Y}_{k-1,n}, X_n)} \\
&- \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)} \Big\} A_{kn}
\end{aligned}$$

Then, in Theorem 2 we analogously define $\tilde{p}(Z)$ as follows:

$$\begin{aligned}
\tilde{p}(Z) \equiv p(X) p^*(S|X; \alpha_s) \prod_{k=1}^{K} \Big\{ & p^*(M_k|S, \overline{A}_{k-1}, \overline{Y}_{k-1}, X; \alpha_m) \\
& \times p(A_k|H_k) p(Y_k|A_k, H_k) \Big\}.
\end{aligned}$$

In this case we constrain the $S$ and $M_k$ models $\forall k$, the rest of the procedure remaining the same. Aside from the form of the identifying functional, the proofs of modified versions of Theorem 1 and Theorem 2 are analogous.

## References


James M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*, pages 189–326, 2004.

Eric J. Tchetgen Tchetgen and Ilya Shpitser. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864, 2014.

Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.