Remark 24. If the actionspace $\mathcal{A}$ is unconstrained, that is consists of all possible assignments of values to variables, (and all actions have equal cost) the optimal action will set the value of all the parents of Y and Algorithm 3 cannot do better than uniform exploration over these arms. In this case, after we use the causal structure to eliminate irrelevent actions prior to taking any samples, the problem of selecting within the remaining actions can be treated as a standard multi-armed bandit problem.

Theorem 25. *Let $\mathcal{A}'$ be the set of all possible assignments of values to the parents of Y. If $\mathcal{A}' \subseteq \mathcal{A}$ and $C(a') \leq C(a)\ \forall(a' \in \mathcal{A}', a \in \mathcal{A}/\mathcal{A}')$ then the optimal action $a^* \in \mathcal{A}'$ and the problem reduces to a standard multi-armed bandit (over actions in in $\mathcal{A}'$).*

Proof. for any action $a \in \mathcal{A}$,

$$\begin{aligned}\mathbb{E}\left[Y|\boldsymbol{X}^c_t, a\right] &= \mathbb{E}_{\mathcal{P}\text{a}_Y \sim \text{P}(\mathcal{P}\text{a}_Y|\boldsymbol{X}^c_t,a)}\left[\mathbb{E}\left[Y|\boldsymbol{X}^c_t, a, \mathcal{P}\text{a}_Y\right]\right] \\ &= \mathbb{E}_{\mathcal{P}\text{a}_Y \sim \text{P}(\mathcal{P}\text{a}_Y|\boldsymbol{X}^c_t,a)}\left[\mathbb{E}\left[Y|\,\mathcal{P}\text{a}_Y\right]\right] \\ &= \mathbb{E}_{\mathcal{P}\text{a}_Y \sim \text{P}(\mathcal{P}\text{a}_Y|\boldsymbol{X}^c_t,a)}\left[\mathbb{E}\left[Y|do(\mathcal{P}\text{a}_Y)\right]\right] \\ &\leq \max_{\mathcal{P}\text{a}_Y} \mathbb{E}\left[Y|do(\mathcal{P}\text{a}_Y)\right] = \mathbb{E}\left[Y|a'\right] \text{ for some } a' \in \mathcal{A}'\end{aligned}$$

This proves the optimal action $a^* \in \mathcal{A}'$

We now consider using importance weighted estimators from Algorithm 3 to estimate the rewards for all actions in $\mathcal{A}'$. The optimal sampling weights η are given by,

$$\eta^* = \arg\min_{\eta} \max_{a \in \mathcal{A}'} \mathbb{E}_a\left[\frac{\text{P}\left\{\mathcal{P}\text{a}_Y(X)|a\right\}}{\sum_{b\in\mathcal{A}} \eta_b \text{P}\left\{\mathcal{P}\text{a}_Y(X)|b\right\}}\right].$$

Note that we now only have to obtain estimates for actions $a \in \mathcal{A}'$, since we know the others to be sub-optimal, so the *max* is only over these actions. However we b still sums over all possible actions in the demoninator of the importance sampling estimator, to allow for the possiblity that playing sub-optimal actions allows more efficient estimation of the optimal actions. We now prove that this is not the case and that, in this specific setting, Algorithm 3 cannot do better than uniform sampling over the actions $a \in \mathcal{A}'$. Now each action $a \in \mathcal{A}'$ consists of a given assignment $\boldsymbol{x}_a$ to $\mathcal{P}\text{a}_Y$.

$$\begin{aligned}a =& do(\mathcal{P}\text{a}_Y = \boldsymbol{x}_a) \implies \mathbb{E}_a\left[\frac{\text{P}\left\{\mathcal{P}\text{a}_Y(X)|a\right\}}{\sum_{b\in\mathcal{A}} \eta_b \text{P}\left\{\mathcal{P}\text{a}_Y(X)|b\right\}}\right] = \frac{1}{\sum_{b\in\mathcal{A}} \eta_b \text{P}\left\{\mathcal{P}\text{a}_Y = \boldsymbol{x}_a|b\right\}} \\ &\implies \eta^* = \arg\max_{\eta}\left[\min_{a\in\mathcal{A}'} \sum_{b\in\mathcal{A}} \eta_b \text{P}\left\{\mathcal{P}\text{a}_Y = \boldsymbol{x}_a|b\right\}\right]\end{aligned}$$

Let N_a denote $\sum_{b\in\mathcal{A}} \eta_b \text{P}\left\{\mathcal{P}\text{a}_Y = \boldsymbol{x}_a|b\right\}$, which can be viewed as an effective number of samples for action a. Choosing $\eta_b = \mathbb{1}\{b \in \mathcal{A}'\}\frac{1}{|\mathcal{A}'|}$, corresponding to uniform exploration over the optimal arms only, yields $N_a = \frac{1}{|\mathcal{A}'|}$ for all a. To do better, we would need to find weights η such that $N_a > \frac{1}{|\mathcal{A}'|}$ for all a. However,

$$\begin{aligned}N_a > \frac{1}{|\mathcal{A}'|}\forall a &\implies \sum_{a\in\mathcal{A}'} N_a > 1 \\ &\implies \sum_{b\in\mathcal{A}} \eta_b \sum_{a\in\mathcal{A}'} \text{P}\left\{\mathcal{P}\text{a}_Y = \boldsymbol{x}_a|b\right\} > 1 \implies \sum_{b\in\mathcal{A}} \eta_b > 1\end{aligned}$$

This violates the fact that η is a distribution over the actions, and thus must have weights that sum to 1, thus completing the proof. □